

This electronic thesis or dissertation has been downloaded from the King's Research Portal at <https://kclpure.kcl.ac.uk/portal/>



Epidemiological study of antibiotic utilization and pneumonia using electronic health records

Sun, Xiaohui

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. <https://creativecommons.org/licenses/by-nc-nd/4.0/>

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Epidemiological Study of Antibiotic Utilization and
Pneumonia Using Electronic Health Records

Xiaohui Sun

Thesis submitted for the degree of Doctor of Philosophy

School of Population Health and Environmental Sciences

King's College London

Abstract

Community acquired pneumonia (CAP) is one of the most common infectious conditions managed in primary care. CAP may complicate simple respiratory tract infections (RTIs) but there is limited evidence available to inform general practitioners (GPs) of the characteristics of RTI patients who may be at the risk of pneumonia. The purpose of this thesis was to analyse the clinical profile of adult RTI patients to identify variables associated with development of pneumonia within 30 days through a prediction modelling study.

The thesis reports a series of inter-related research studies that analysed electronic health records data from the Clinical Practice Research Datalink (CPRD). The research addressed four inter-related objectives. The wider antimicrobial stewardship context was explored through an analysis of antibiotic prescribing records from English general practices participating CPRD from 2014 to 2017. Antibiotic prescriptions for the main groups of common infections managed in the community including respiratory infections, genito-urinary infections (GUTI), infectious skin conditions, eye infections were evaluated. An annual relative reduction rate (RRR) of 6.9% for total antibiotic prescription was detected during the four-year period in the English primary care. Respiratory conditions remained to be the most frequent indications for antibiotic prescriptions among informatively coded consultations, also showed the greatest reduction in prescription rates.

Next, secular trends in the incidence of clinically-diagnosed CAP, clinically-suspected CAP, influenza and pleural infections were evaluated using CPRD data from 2002 to 2017. Clinically-diagnosed CAP incidence was found to increase over time with an accelerated trend after 2010. For clinically-suspected CAP, an overall contemporaneous trend with an average increasing rate being 3.8% from 2002 to 2008 whereas a faster decline rate of 4.9% thereafter until 2017. Study results together with previous research findings suggested that antibiotic prescribing practice and clinically coding behaviour partly contributed to the apparent increase in clinically-diagnosed CAP in primary care settings.

A systematic review of current evidence of prognostic factors for CAP was conducted to identify candidate predictors for the prediction modelling study. 33

prognostic factors for CAP were identified which could be categorized into six groups: patients' demographic characteristics, lifestyle, environmental exposures, health conditions, medication prescriptions, disease prevention interventions, clinical management and clinical investigations.

Based on previous study findings, prediction modelling study was conducted with an inclusive approach for possible candidate predictors generated from CPRD data from 2002 to 2017. Analysis included 108,842 patients who consulted for RTIs of whom 16,289 patients re-consulted with pneumonia within 30 days after the RTI index date. Data were analysed using machine learning algorithms for variable selection. Variable selection employed and compared random forest, simple logistic regression, and penalized regression models (Lasso, Ridge and Elastic net). Prediction models were developed using the classification and regression tree (CART) approach, as well as simple logistic regression. Internal and temporal validation were performed. Older age, comorbidity and initial presentation with lower respiratory tract infections (LRTIs) were identified as the main predictors of pneumonia diagnosis. Among patients presented with LRTIs, patients older than 85 remained at higher risk of pneumonia re-consultation despite antibiotic prescriptions were offered; those age between 76 and 85 with two or more comorbidities risk of pneumonia re-consultation persisted even if antibiotic prescriptions were issued. LRTI patients younger than 65 without asthma drug or immunosuppressants treatments appeared to have higher risk of pneumonia re-consultation if clinical discretion did not lead to antibiotic treatment. However, cautions are needed when interpreting such counter-intuitive findings as allocation to antibiotic treatment as well as other disease management procedures were not randomized and confounding by disease indications. Therefore, disease pattern identified among LRTI patients indicated that more attention should be paid to subgroup of LRTI patients to investigate the underlying reasons of primary onset of clinical conditions.

Machine learning techniques may allow the identification of novel disease pattern comparing to conventional modelling approaches, which could be deployed to generate research hypothesis, individualized research designs for inventory clinical trials or provide insights for health policy development.

Table of Contents

Abstract	2
Table of Contents	4
List of tables	7
List of figures	10
List of tables in appendix.....	13
List of figures in appendix	15
List of abbreviations	17
Acknowledgement	21
Chapter One : Introduction	22
<i>1.1 The changing role of pneumonia</i>	<i>22</i>
<i>1.2 Epidemiology and microbiology of community acquired pneumonia (CAP)</i>	<i>23</i>
<i>1.3 Clinical presentation of community acquired pneumonia, differential diagnosis and complications</i>	<i>25</i>
<i>1.4 Antibiotic and respiratory tract infections (RTIs) among the adult population in the UK primary care settings.....</i>	<i>27</i>
<i>1.5 Current challenges of RTI management in primary care</i>	<i>28</i>
<i>1.6 Rationale for the thesis</i>	<i>31</i>
Chapter Two : Thesis research question, aim, objectives and scope.....	35
<i>2.1 Research question</i>	<i>35</i>
<i>2.2 Research aim</i>	<i>35</i>
<i>2.3 Research objectives</i>	<i>35</i>
<i>2.4 Research scope</i>	<i>36</i>
<i>2.5 Structure of the thesis</i>	<i>37</i>
Chapter Three : Understanding electronic health records (EHRs) and clinical practice research datalink (CPRD)	38
<i>3.1 Overview of electronic health records (EHRs).....</i>	<i>38</i>
<i>3.2 EHRs in medical research</i>	<i>39</i>

3.3 EHRs in UK primary care research	46
3.4 The clinical practice research datalink (CPRD) Gold	49
3.5 Recoding of RTI, chest infection, pneumonia and pleural infection.....	53
Chapter Four : Reducing antibiotic prescribing in primary care in England from 2014 to 2017: a population-based cohort study.....	58
4.1 Introduction	58
4.2 Methodology	60
4.3 Results	64
4.4 Discussion.....	72
4.5 Conclusion and implication for future research.....	76
Chapter Five : Pneumonia incidence trends in UK primary care from 2002- 2017: a population-based cohort study	78
5.1 Introduction	78
5.2 Methodology	80
5.3 Results	83
5.4 Discussion.....	89
5.5 Conclusion and implication for future research.....	93
Chapter Six : Prognostic factors for development of a model in community- acquired pneumonia in adults: a systematic review	94
6.1 Introduction	94
6.2 Objectives	97
6.3 Methods.....	97
6.4 Results	100
6.5 Discussion.....	137
6.6 Conclusions and implications for further research.....	139
Chapter Seven : Clinical prediction model development for adult RTI patients who reconsulted with pneumonia in 30 days (Introduction and methodology).....	140
7.1 Introduction	140

7.2 Methodology	159
7.3 Data governance approval	190
Chapter Eight : Clinical prediction model development for adult RTI patients who reconsulted with pneumonia within 30 days (Results, Discussion and Conclusion)	191
8.1 Results	191
8.2 Discussion.....	253
8.3 Conclusion and future implications	263
Chapter Nine : Overall discussion and conclusions	264
9.1 Summary of main findings	264
9.2 General discussion and reflections on the thesis	266
9.3 Strengths and limitations of this thesis	279
9.4 Conclusions and future implications	280
References	281
Appendices:.....	333
Appendix A: Comparisons of clinical terminology systems	333
Appendix B: Medical codes associated with antibiotic prescriptions in primary care in England	339
Appendix C: Code lists for antibiotic prescription study	343
Appendix D: Searching strategies for systematic review	384
Appendix E: Excluded studies by full text.....	388
Appendix F: Interim results during model development	399
Appendix G: TRIPOD Checklist for reporting prediction modelling study	456
Appendix H: Publications and outputs during thesis	458

List of tables

Table 1.1: National Institute for Health and Care Excellence (NICE) guideline recommendations for immediate antibiotic treatment or investigation for RTI patients (NICE, 2008a)	30
Table 3.1: Main data files of CPRD (CPRD, 2017).....	52
Table 4.1: Numbers of general practices contributing to CPRD from 2014 to 2017. Figures are mid-year counts.	62
Table 4.2: Numbers of antibiotic prescriptions, and antibiotic prescribing rates, by year. Figures are frequencies except where indicated.....	67
Table 4.3: Distribution of antibiotic prescriptions by broad groups of indications. Figures are frequencies except where indicated.....	68
Table 5.1: Number of incidence events of pneumonia and related conditions. Figures are frequencies except where indicated.....	85
Table 5.2: Joinpoint regression estimates for annual percent change (APC).	86
Table 6.1: Characteristics of included studies.....	104
Table 6.2: Characteristics of included studies (continued)	111
Table 6.3: Individual predictors identified from systematic review	122
Table 6.4: Individual predictors identified from systematic review (continued).....	126
Table 6.5: Results of Quality in Prognosis Studies (QUIPS) assessment for prognostic factors studies for CAP	130
Table 6.6: Prognostic effects of predictors identified from low bias primary studies	134
Table 7.1: Illustration of common ensemble methods	150
Table 7.2: Gini Index, Chi-Square and Information gain as splitting criteria for classification tree.....	174
Table 8.1: Descriptive statistics of demographic characteristics for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.....	193
Table 8.2: Descriptive statistics for frailty for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.	194

Table 8.3: Descriptive statistics of chronic conditions including co-morbidity for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.....	195
Table 8.4: Descriptive statistics of initial RTIs for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.	196
Table 8.5: Descriptive statistics of medical management for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.	197
Table 8.6: Top 15 variables as selected by simple logistic regression, penalized regression and random forest (full model)	204
Table 8.7: Classification and regression tree (CART) model performance for internal and temporal validations	208
Table 8.8: Simple logistic model using same variables from those for CART (full model)	212
Table 8.9: Simple logistic model performance for internal and temporal validations (full model).....	213
Table 8.10: Top 15 variables as selected by simple logistic regression, penalized regression and random forest for lower respiratory tract infection (LRTI) patients	217
Table 8.11: CART model performance for internal and temporal validations (LRTI model)	220
Table 8.12: Simple logistic model using same variables from those for CART (LRTI model)	224
Table 8.13: Simple logistic model performance for internal and temporal validations (LRTI model)	225
Table 8.14: Top 10 variables as selected by simple logistic regression, penalized regression and random forest for upper respiratory tract infection (URTI) model..	229
Table 8.15: CART model performance for internal and temporal validations (URTI model)	232
Table 8.16: Simple logistic model using same variables from those for CART (URTI model)	236
Table 8.17: Simple logistic model performance for internal and temporal validations (URTI model).....	237

Table 8.18: Summary of discriminative performances of developed models as measured by AUROC	240
Table 8.19: Top 15 variables as selected by simple logistic regression, penalized regression and random forest for sensitivity analysis of full model	242
Table 8.20: CART model performance for internal and temporal validations (sensitivity analysis of full model)	245
Table 8.21: Simple logistic model using same variables from those for CART (sensitivity analysis of full model)	249
Table 8.22: Simple logistic model performance for internal and temporal validations (sensitivity analysis of full model)	250

List of figures

Figure 1.1: Illustration of major RTIs included in this thesis	33
Figure 3.1: RTIs and pleural infections recoding process in CPRD	57
Figure 4.1: Proportion of patients prescribed antibiotics in year by age-group and calendar year.	65
Figure 4.2: Forest plot showing annual relative reduction (95% confidence interval) in antibiotic prescribing for all antibiotics and broad-spectrum β -lactam antibiotics between 2014 and 2017 for sub-groups of age and gender and different prescribing indications. Estimate were adjusted for age, gender and clustering by practice.....	70
Figure 4.3: Bar chart showing changes from 2014 to 2017 in the proportion of antibiotic prescriptions for different antibiotic classes for males and females.	72
Figure 5.1: Trends in pneumonia and related conditions for both men (blue) and women (red) 2002-2017. Rates are per 1,000 person-years.....	87
Figure 5.2: Age-specific rates for clinically-diagnosed pneumonia and clinically-suspected pneumonia for males (blue) and females (red). Rates are per 1,000 person-years	88
Figure 6.1: PRISMA flow diagram outlining systematic review process.....	101
Figure 7.1: Flow chart of pneumonia case ascertainment.....	161
Figure 7.2: Flow charts of non-pneumonia case sample selection based on pre-defined stratified sampling criteria.....	163
Figure 7.3: Exemplar of decision tree	172
Figure 7.4: Illustration of decision tree splitting on sex vs travel class to predict survival status of adult passengers using <i>Titanic</i> data.	173
Figure 7.5: Illustration of bagging process	177
Figure 7.6: Illustration of random forest for classification trees (Carriquiry et al., 2019).	178
Figure 7.7: Common regression models based on key metrics.....	180
Figure 7.8: Illustration of under fit vs over fit model	182
Figure 7.9: Trade-off between model variance and bias (Huilgol, 2020).....	183
Figure 7.10: Illustration of the optimal regression model.....	184
Figure 7.11: Lambda for three penalize regressions (full model).....	187
Figure 7.12: Exemplar of difference between ridge and lasso regression (Efron and Hastie, 2016b)	188

Figure 8.1: Variable importance results by random forest models with Mtry 5- 30 (5 increments) for full model.....	199
Figure 8.2: Cross validation curve of lambda for lasso regression	201
Figure 8.3: Variable importance for full model based on simple logistic regression and penalized regressions	202
Figure 8.4: Classification tree based on variable selection results for study population (full model).....	207
Figure 8.5: Receiver operating characteristic (ROC) curve for internal and temporal validation of CART model (full model).....	209
Figure 8.6: Calibration plots for internal and temporal validation of CART model (full model).....	210
Figure 8.7: ROC curve for internal and temporal validation of simple logistic model (full model).....	214
Figure 8.8: Calibration plots for internal and temporal validation of simple logistic regression (full model)	215
Figure 8.9: CART model for patients presented with LRTIs.....	219
Figure 8.10: ROC curve for internal and temporal validation of CART model (LRTI model)	221
Figure 8.11: Calibration plots for internal and temporal validation of CART (LRTI model)	222
Figure 8.12: ROC curve for internal and temporal validation of simple logistic model (LRTI model)	226
Figure 8.13: Calibration plots for internal and temporal validation of simple logistic regression (LRTI model).....	227
Figure 8.14: CART model for patients presented with URTIs	230
Figure 8.15: ROC curve for internal and temporal validation of CART model (URTI model)	233
Figure 8.16: Calibration plots for internal and temporal validation of CART (URTI model)	234
Figure 8.17: ROC curve for internal and temporal validation of simple logistic model (URTI model).....	238
Figure 8.18: Calibration plots for internal and temporal validation of simple logistic regression (URTI model)	239

Figure 8.19: CART model sensitivity analysis of full model	244
Figure 8.20: ROC curve for internal and temporal validation of CART model (sensitivity analysis of full model)	246
Figure 8.21: Calibration plots for internal and temporal validation of CART model (sensitivity analysis of full model)	247
Figure 8.22: ROC curve for internal and temporal validation of simple logistic regression model (sensitivity analysis of full model)	251
Figure 8.23: Calibration plots for internal and temporal validation of simple logistic regression model (sensitivity analysis of full model)	252

List of tables in appendix

Table A 1: Comparisons within primary care terminology systems (Read code version 2 and International Classification for Primary Care, ICPC-2)	333
Table A 2: Comparison of main structure between Read Code-2 and International Classification of Disease, ICD-10	335
Table A 3: Comparison of main structure between ICPC-2 and ICD-10	337
Table A 4: Medical codes associated with 99.8% antibiotic prescriptions in CPRD ranked descendent according to frequencies	339
Table A 5: Read codes for respiratory conditions.....	343
Table A 6: Read codes for genitourinary conditions.	356
Table A 7: Read codes for skin conditions.	368
Table A 8: Read codes for eye conditions.	378
Table A 9: Excluded studies by full text	389
Table A 10: General health status of study cohort frailty vs comorbidity	399
Table A 11: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for full model	400
Table A 12: Lambda and alpha for elastic net (full model)	402
Table A 13: Absolute value of coefficients of three penalized regression models (full model)	406
Table A 14: Variable importance ranking based on top 20 variables across three penalized regression models (full model)	408
Table A 15: Machine learning model comparison using full dataset with 10-fold cross validation.....	409
Table A 16: Machine learning model using full dataset with 10-fold cross validation ROC curves	410
Table A 17: Comparison statistics of model variables for development data and temporal validation data (full model). Figures are frequencies (column percentages) except where indicated.....	414
Table A 18: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for LRTI model.....	417
Table A 19: Lambda and alpha for elastic net (LRTI model).....	419
Table A 20: Absolute value of coefficients of three penalized regression models (LRTI model)	423

Table A 21: Variable importance ranking based on top 18 variables across three penalized regression models (LRTI model)	426
Table A 22: Comparison statistics of model variables for development data and temporal validation data (LRTI model). Figures are frequencies (column percentages) except where indicated.....	428
Table A 23: Variable importance ranking based on top 10 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for URTI model	431
Table A 24: Lambda and alpha for elastic net (URTI model)	432
Table A 25: Absolute value of coefficients of three penalized regression models (URTI model).....	436
Table A 26: Variable importance ranking based on top 15 variables across three penalized regression models (URTI model)	439
Table A 27: Comparison statistics of model variables for development data and temporal validation data (URTI model). Figures are frequencies (column percentages) except where indicated.....	441
Table A 28: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for sensitivity analysis model	444
Table A 29: Lambda and alpha for elastic net (sensitivity analysis of full model) .	446
Table A 30: Absolute value of coefficients of three penalized regression models (sensitivity analysis of full model).....	450
Table A 31: Variable importance ranking based on top 20 variables across three penalized regression models (sensitivity analysis of full model).....	453
Table A 32: Comparison statistics of model variables for development data and temporal validation data (sensitivity analysis of full model). Figures are frequencies (column percentages) except where indicated.	455
Table A 33: TRIPOD checklist for prediction model development and validation.	456

List of figures in appendix

Figure A 1: Number of trees for random forest 50, 100 and 150 for full model	399
Figure A 2: Lambada for 3 penalize regressions (full model).....	401
Figure A 3: Alpha for elastic net model against misclassification error (full model)	405
Figure A 4: Tuning parameter for CART full model	413
Figure A 5: Number of trees for random forest 50, 100 and 150 for LRTI model ..	415
Figure A 6: Variable importance results by random forest models with Mtry 5- 30 (5 increments) for LRTI model	416
Figure A 7: Lambada for 3 penalize regressions (LRTI model).....	418
Figure A 8: Alpha for elastic net model against misclassification error (LRTI model)	422
Figure A 9: Variable importance for LRTI model based on simple logistic regression and penalized regressions.....	425
Figure A 10: Tuning parameter for CART LRTI model.....	427
Figure A 11: Number of trees for random forest 50, 100 and 150 for URTI model	429
Figure A 12: Variable importance results by random forest models with Mtry 5- 30 (5 increments) for URTI model.....	430
Figure A 13: Lambada for 3 penalize regressions (URTI model)	431
Figure A 14: Alpha for elastic net model against misclassification error (URTI model)	435
Figure A 15: Variable importance for URTI model based on simple logistic regression and penalized regressions	438
Figure A 16: Tuning parameter for CART URTI model	440
Figure A 17: Number of trees for random forest 50, 100 and 150 for sensitivity analysis of full model	442
Figure A 18: Variable importance results by random forest models with Mtry 5- 30 (5 increments) for sensitivity analysis of full model.....	443
Figure A 19: Lambada for 3 penalize regressions (sensitivity analysis model of full model)	445
Figure A 20: Alpha for elastic net model against misclassification error (sensitivity analysis of full model).....	449

Figure A 21: Tuning parameter for CART model (sensitivity analysis of full model)

..... 454

List of abbreviations

ACE inhibitors	Angiotensin-converting enzyme inhibitors
ACSC	Ambulatory care sensitive conditions
AIC	Akaike Information Criterion
AIDS	Acquired immune deficiency syndrome
AMR	Antimicrobial resistance
APACHE	Acute Physiology and Chronic Health Evaluation
AAPC	Average annual percent changes
APC	Annual percent changes
ARVC	Arrhythmogenic right ventricular cardiomyopathy
ASA	American Society of Anaesthesiologists
ASR	Age standardised rate
ATS	American Thoracic Society
BMI	Body mass index
BMA	British Medical Association
BNF	British National Formulary
BOOP	Bronchiolitis obliterans organizing pneumonia
BTS	British Thoracic Society
CAP	Community-acquired pneumonia
CART	Classification and regression tree
CCBs	Calcium channel blockers
CDC	Centers for Disease Control and Prevention
CENTRAL	Cochrane Database of Systematic Reviews and Cochrane Central Register of Controlled Trials
CGA	Comprehensive Geriatric Assessment
CHD	Coronary heart disease
CHI	Community Health Index
CKD	Chronic kidney disease
CLD	Chronic liver disease
CND	Chronic neurological disease
CNS	Central nervous system
COPD	Chronic obstructive lung disease
Covid-19	Coronavirus 2019

CPRD	Clinical Practice Research Datalink
CRB-65	Confusion, raised respiratory rate, low blood pressure and age 65 and above
CRD	Chronic respiratory disease
CRE	Carbapenem-resistant Enterobacteriaceae
CRP	C-reactive protein
CVD	Cardiovascular disease
DCA	Decision curve analysis
DM	Diabetes mellitus
eFI	Electronic frailty index
EHR	Electronic health record
EPV	Event per variable
ESPAUR	The English Surveillance Programme for Antimicrobial Utilisation and Resistance
GBD	Global Burden of Disease
GCS	Glasgow Coma Scale
GP	General practitioner
GUTI	Genito-urinary tract infection
HES	Health Episode Statistics
Hib	Haemophilus influenza type b
HIE	Health information exchange
H-L test	Hosmer–Lemeshow test
HP	Helicobacter pylori
ICD	International Classification of Diseases
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
ICPC	International Classification of Primary Care
ICU	Intensive care unit
IMD	Index of Multiple Deprivation
ISAC	Independent Scientific Advisory Committee
ISO	International Organization for Standardization
LSOA	Lower Layer Super Output Area
LRTI	Lower respiratory tract infection

MAR	Missing at random
MCAR	Missing completely at random
MHDS	Mental Health Dataset
MI	Myocardial infarction
MHRA	Medicines and Healthcare Products Regulatory Agency
MNAR	Missing not at random
MRSA	Methicillin-resistant Staphylococcus aureus
MSA	Measurement system analysis
MSE	Mean squared error
Mtry	Number of variables randomly sampled when creating the tree
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIH	National Institutes of Health
Ntree	Number of trees
OADs	Oral antidiabetic agents
ONS	Office for National Statistics
OOB	Out of bag
OXMIS	Oxford Medical Information System
PACU	Post anaesthesia recovery unit
PCA	Principle component analysis
PCT	Procalcitonin
PCV7	Pneumococcal conjugate 7 vaccine
PCV13	13-valent pneumococcal conjugate vaccine
PERRLA	Pupils equal, round, and reactive to light and accommodation
PHE	Public Health England
PPIs	Proton pump inhibitors
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analysis
PROGRESS	The Prognosis research strategy
PTE	Pulmonary thromboembolism
QOF	Quality Outcome Framework
QUIPS	Quality in Prognosis Studies

RECORD	REporting of studies Conducted using Observational Routinely-collected health Data
ROC	Receiver Operating Characteristics
RRR	Relative reduction rate
RTI	Respiratory tract infection
SE	Standard error
SICU	Surgical intensive care unit
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SNP	Single nucleotide polymorphisms
T2DM	Type 2 diabetes
TB	Tuberculosis
THIN	Health Improvement Network
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
URTI	Upper respiratory tract infection
UK	United Kingdom
US	United States
UTS	Up-to-standard
VAMP	Value Added Medical Products
WBC	White blood cell
WHO	World Health Organization

Acknowledgement

I would like to express my deep gratitude to my first supervisor Professor Martin Gulliford who supervised both my master dissertation and PhD thesis. It is my unique opportunity to be mentored by him. Working with him is a great experience to learn from someone with professional expertise, great intelligence and insightfulness. My appreciation also goes to my second supervisor Dr Abdel Douiri whose excellent advice made me dare to push the boundaries of my own work. I would also like to thank Professor Christopher McKevitt as well as my thesis progression committee: Dr Mark Ashworth, Professor Judith Green, Dr Patrick White.

I am also very grateful to Professor Xiaodong Yang and Professor Zong'an Liang from West China clinical medical school who ignited my passion for respiratory medicine. I am always thankful for all the patients I served for, from whom I have learned a lot. I would also acknowledge all researchers and patients who have contributed to CPRD which allows our research to be built upon.

My huge thanks are to my family and friends who have always been the source of constant support all the way through. To my dearest Michelle, my pride and joy, having you in my life always gives me the willingness to be a better person for you, for us.

This thesis is completed in the memorial of my beloved father, a sincere man, a dedicated researcher, who is always there giving me spiritual guidance. Daddy, I love you...

For the things we believed in:

天行健，君子以自强不息；地势坤，君子以厚德载物。

Chapter One : Introduction

1.1 The changing role of pneumonia

Since 2009, November 12th has been marked as World Pneumonia Day, with the aim of increasing awareness that pneumonia continues to be a global clinical and public health concern (WHO, 2018b). Despite the availability of effective treatments and prevention interventions, pneumonia claims more young lives than other infectious diseases, and is the leading infectious killer of children under 5-years old (WHO, 2016b). According to the Global Burden of Disease (GBD), Injury, and Risk Factors Study 2015, lower respiratory tract infections (LRTIs) were responsible for more than 2.7 million deaths and affected 291.8 million people of all ages worldwide (GBD 2015 LRI Collaborators, 2017). Pneumococcal pneumonia is by far the single largest cause of death among all LRTIs, accounting for 55.4% of LRTI mortality (Aliberti et al., 2019, GBD 2015 LRI Collaborators, 2017). However, even this overwhelming death toll does not arouse public attention (Wardlaw et al., 2006); on the contrary, pneumonia has been overshadowed as a priority on the global health agenda.

The misalignment between the perceived and actual severity of pneumonia represents a triumph of the antibiotic era, which has exerted a monumental impact in medical history as well as influencing the disease profile of pneumonia. The initial fear of pneumonia was derived from its case fatality rate of 30% to 40%; pneumonia was referred to as the ‘captain of the men of death’ during the pre-antibiotic era (Podolsky, 2006). After the introduction of the first sulphonamide drug during the 1930s and more importantly penicillin in the 1940s, pneumonia, which might previously have been life threatening, could be treated effectively with antibiotics. The discovery and synthesis of further new antibiotics enriched the pipeline of anti-microbial treatment resulting in a great difference in the implications of pneumonia for medical professionals and the general public between the pre and post antibiotic eras (Aliberti et al., 2019). The stark contrast has made the public perception of the possible consequences of pneumonia less concerning. It may also have been reflected in the undesirably low uptake of influenza and pneumococcal vaccinations as main

prevention strategies for pneumonia (Örtqvist, 2001, Lim et al., 2001a). In the field of medical research, funding for pneumonia research remains at a low level relative to the burden of disease. For example, funding from the National Institutes of Health (NIH) for pneumonia research was 12% of that for acquired immune deficiency syndrome (AIDS) even though the incidence of pneumonia is 30 times higher (Lim, 2015, Head et al., 2014). Although research investment may not be the only driving force in science innovation, it appears that no important advances in terms of clinical diagnosis, treatment or medical evidence for pneumonia have been observed during the past two decades (Woodhead et al., 2011a, Mandell et al., 2007). For instance, there were no substantial updates noted in the key recommendations in the British Thoracic Society (BTS) guidelines for adult pneumonia management between 2009 and 2015 (Lim et al., 2015).

The perception that pneumonia is not a severe condition may result from other advances in medical technology. More severe cases, that would not have been successfully treated in the past, can now be managed effectively. Pneumonia may sometimes have been relegated from principal to secondary or accompanying diagnosis in hospital records. This has resulted in a reduction in hospitalization rates with pneumonia being the primary reason for admission but with a contemporaneous increase when pneumonia was grouped into a related principal diagnosis such as respiratory failure or sepsis (Lindenauer et al., 2012). Also, recent evidence has shown that there has been no obvious decrease in the pneumonia fatality rate since the 1950s (Aliberti et al., 2019).

1.2 Epidemiology and microbiology of community acquired pneumonia (CAP)

CAP is one of the most common causes of morbidity and mortality globally and regionally (Mandell et al., 2007, Musher and Thorner, 2014). Pneumonia incidences has been estimated to lie between 1.5 and 14 per 1000 person years depending on the region, season and demographic variables of the population (Millett et al., 2013, Ochoa-Gondar et al., 2008, File Jr and Marrie, 2010). Medical professionals ranging from general practitioners (GPs) to specialists may encounter pneumonia at some

point. For example, neurologists have to treat aspiration pneumonia among stroke patients and orthopaedic surgeons manage hypostatic pneumonia for hip fracture patients (Weingarten et al., 2002). In most healthcare systems, CAP is a frequent cause of emergency hospital visits and admission for hospital management (Schappert and Burt, 2006, Weiss et al., 2006). In the USA for instance, hospital stays due to CAP were only surpassed by livebirth (Pfundner et al., 2006) with recent annual incidence of hospital admission relating to pneumonia among the adult population reported to be more than 24.8 cases per 10,000. The highest of these rates were observed among the older population (63.0 per 10,000 adults for 65 to 79 age group and 164.3 per 10,000 adults for those 80 and above) (Jain et al., 2015a). Recently increasing trends for CAP hospitalization have been noticed in the UK, especially among the older population (Quan et al., 2016c, Trotter et al., 2008).

The concept of CAP was initially introduced by the American Thoracic Society (ATS) in 1993, emphasising the pneumonia acquisition environment and linking it directly to possible causative microbial organisms to guide initial empirical antimicrobial treatment (Ewig et al., 2010, Falcone et al., 2011, Niederman, 1998, ATS, 1993). More than just a classification, CAP represents a clinical concept that provides insight into an antibiotic treatment plan which would cover the majority of pneumonia episodes in community settings. Although most CAP patients respond well to standard empirical management, researchers have noted the heterogeneity and changing patterns of CAP aetiology (Torres et al., 2014, Prina et al., 2015).

Streptococcus pneumoniae (*S. pneumoniae*) has been confirmed as the predominant pathogen responsible for CAP regionally and globally across all age groups usually followed by *Haemophilus influenzae* (Lim et al., 2001b, Welte et al., 2012, Drijkoningen and Rohde, 2014, Howard et al., 2005). However, a trend towards declining *S. pneumoniae* incidence was noticed with the increased isolation of respiratory viruses and, antibiotic resistant serotypes like *Methicillin-resistant Staphylococcus aureus* (MRSA) as well as atypical bacteria such as *Mycoplasma*, *Chlamydia* and *Legionella spp* (Ruuskanen et al., 2011, Arnold et al., 2007, Principi and Esposito, 2012). The proportion of CAP cases caused by atypical bacterial pathogens was estimated to be 22% and is closely associated with high mortality (Arnold et al., 2007, Prina et al., 2015). The evolving CAP causal pathogen profile

has implications for non-typical or latent clinical presentations, even the need for a refined treatment plan among subgroup of patients.

1.3 Clinical presentation of community acquired pneumonia, differential diagnosis and complications

CAP may be considered as a broad term for a group of common disorders with great variation in clinical manifestations ranging from mild pneumonia presenting with signs of LRTI without other obvious causes to severe cases that are potentially life threatening. A definite diagnosis of CAP is established based on clinically suspected pneumonia commonly presenting with acute respiratory infection symptoms and new infiltrate confirmed via chest x-rays (Prina et al., 2015, Mandell et al., 2007, Lim et al., 2009a, Eccles et al., 2014). However, a pneumonia diagnosis is not always straightforward. The clinical and radiological findings contributing to pneumonia diagnosis may be inconsistent, low inter-observer agreement was reported among radiologists as final arbiters in terms of CAP ascertainment (Albaum et al., 1996, Hopstaken et al., 2004). Importantly, X-rays are often not readily available in primary care settings (Cherryman, 2006, ESR and WONCA, 2010); pneumonia diagnosis in the community largely relies on clinical symptoms, medical history and thorough physical evaluation. Classic pneumonia symptoms include cough, dyspnoea, sputum production, fever and positive physical findings of pulmonary consolidation (Musher and Thorner, 2014, Prina et al., 2015, Lim et al., 2009a, Eccles et al., 2014). In subgroups such as older patients, less typical symptoms could be presented such as confusion, loss of appetite and even absence of fever. For some atypical pathogens, extrapulmonary symptoms are generally the initial onset of pneumonia. Pneumonia caused by *Mycoplasma pneumoniae* can present with otitis, pharyngitis, skin disorders (Stevens-Johnson like syndrome) and even anaemia (Cunha, 2006).

The differential diagnosis of pneumonia varies from common respiratory infectious conditions to non-infectious diseases as many diseases have clinical manifestations that mimic pneumonia (Metlay and Fine, 2003, Maskell, 2010). Common pneumonia symptoms are non-specific but are shared with many respiratory and non-respiratory

conditions. Generally, for mild to moderate CAP, the main differential diagnosis is upper respiratory tract infections (URTIs), whereas for severe CAP, the pneumonia diagnosis in itself is less complicated but lies in the differentiation of other life-threatening conditions such as pulmonary embolism and heart failure (Mangini et al., 2013, Musher and Thorner, 2014). In patients with repeated onset of pneumonia or non-resolved pneumonia within 6 weeks for example, alternative causes should be considered including lung cancer, malignancy, non-infectious pneumonitis or other underlying immunological conditions (Kuru and Lynch III, 1999, Prina et al., 2015).

Most CAP patients respond to appropriate antibiotic treatment (Niederman et al., 2001, McCabe et al., 2009), but some progress to undesirable clinical outcomes or death. The reasons for complications arising from the initial infection such as empyema, lung abscess, endocarditis, sepsis or respiratory failure could be multifactorial. In some cases, a weakened immune system could contribute to the onset of infectious complications despite appropriate antibiotic treatment; the initial empirical antibiotic regimen does not cover atypical or drug resistant causal pathogens. In other cases, complications simply arise from incorrect or delayed diagnosis (Wunderink and Waterer, 2017, Prina et al., 2015). Apart from infectious or respiratory complications, both acute and long-term cardiac events as common comorbid complications have been observed to have close associations with pneumonia. These include myocardial infarction (MI), cardiac arrhythmias, congestive heart failure, pulmonary embolism, and stroke (Corrales-Medina et al., 2012, Corrales-Medina et al., 2013, Violi et al., 2017, Eurich et al., 2017). Such increased recognition is shifting the view of CAP from incident infectious disease confined to the pulmonary system to a systematic process involving multiple extrapulmonary adverse health consequences. This has implications for the CAP care bundle as well as informing cautious antibiotic selection. For example, macrolides have been reported to increase the risk of cardiac arrhythmia and cardiac arrest (Schembri et al., 2013, Mortensen et al., 2014). In addition, it has provided more comprehensive perspectives for medical research in pneumonia.

1.4 Antibiotic and respiratory tract infections (RTIs) among the adult population in the UK primary care settings

RTIs represent some of the most common conditions managed by GPs in the UK (Ashworth et al., 2005). Empirical antibiotic treatment is a cornerstone of RTI management in primary care especially when bacterial infection is likely, or patients are at high risks of developing infectious complications (NICE, 2008a). In UK primary care settings, RTIs among adults are generally classified into three main groups for clinical management purpose: self-limiting RTIs, chest infection and pneumonia (NICE, 2019c, NICE, 2014, NICE, 2015, NICE, 2008a). These are not distinct definitions based on aetiology, rather represents classifications developed with different emphasis on antibiotic prescribing strategies and overlapping clinical concepts.

Self-limiting RTIs, also referred to as uncomplicated respiratory infections have a predominantly viral aetiology for which an antibiotic therapy is not generally indicated. Common self-limiting RTIs, including acute otitis media, acute sore throat/acute pharyngitis/acute tonsillitis, common cold, acute rhinosinusitis together with acute cough/acute bronchitis are grouped into this category. Delayed or immediate antibiotic prescription is only recommended for patients with suspected purulent infections or at risk of developing severe infectious complications including pneumonia, mastoiditis, peritonsillar abscess, peritonsillar cellulitis, intraorbital or intracranial complications (NICE, 2008a). Pneumonia is the most common severe complication from an RTI (Gulliford et al., 2016) and delayed antibiotic treatment was found to increase the risk of death (Metlay and Fine, 2003).

Chest infection mainly comprises two scenarios: acute bronchitis and CAP. For acute bronchitis, antibiotic treatment is generally not indicated unless patients are systematically unwell or have increased risk of complications (NICE, 2015). If clinical assessment leads to a pneumonia diagnosis, then clinical management follows the recommendations for CAP (NICE, 2015, NICE, 2014). For acute cough, which is also a frequent respiratory reason for antibiotic prescription (Gulliford et al., 2014a), the prescribing strategy for adult patients follows clinical guidelines either

for self-limiting RTIs or CAP if the acute cough is considered to be associated with an RTI (NICE, 2019b).

1.5 Current challenges of RTI management in primary care

During the initial clinical management of RTI patients in primary care, clinical discretion involves answering diagnostic as well as prognostic questions. Each question corresponds to a specific clinical decision-making process. If clinical findings, medical history and physical examination indicate that RTI should be considered for medical management, the first diagnostic question ‘Is it of bacterial origin?’ informs the decision on whether antibiotics should be prescribed. If bacterial infection is suggested, the next question would be ‘Which antibiotic(s) should be given to the patients or place of treatment (community vs hospital settings) based on an assessment of severity?’ Even if a concrete infectious diagnosis is not well established, for patients at high risk of complications, the prognostic question ‘Is the patient more likely to develop complications that is antibiotic responsive?’ corresponds to antibiotic prescribing strategy at the point of care. All of these decisions emphasize the importance of timely correct diagnosis and accurate prognosis leading to prompt treatment and to the holding back of unnecessary medication which eventually result in better health outcomes.

In primary care settings, the concrete differentiation between infectious and non-infectious conditions is usually challenging as near patient tests like C-reactive protein (CRP) testing are not routinely performed in UK primary care settings (Cooke et al., 2015, Huddy et al., 2016). If clinical symptoms strongly indicate infection, effective treatment for infectious conditions based on the determination of probable causes may be beneficial. However, antibiotic treatment for most RTI s in primary care settings remain empirical (Wunderink and Waterer, 2017, Lim et al., 2009a). This raises another concern: does one remedy fits all? The apparent answer is no, and even when patients respond well to the same treatment plan, they may not necessarily share the same cause of disease. Given that current microbiological tests do not allow rapid determination of causal pathogens in most clinical settings, it has been suggested that underlying reasons responsible for the onset of RTI symptoms

are usually not fully investigated (Metlay and Fine, 2003). Therefore, additional information on the possible causes or patient characteristics that may be associated with unfavourable health events still contribute to better treatment decisions. This is consistent with clinical guideline recommendations that proactive antibiotic treatment or clinical investigations should be given to RTI patients at higher risks of complications (NICE, 2019b, NICE, 2008a).

However, the risk factors mentioned in guidelines Table 1.1 may not be practical for routine care. For example, the first point ‘systemically very unwell’ is not tangible as clinicians with different training or practicing backgrounds may not always reach to consensus about the general ill health status of the same patients. Also, for the second point, if the patient’s symptoms already suggest severe infectious complications and would eventually result in antibiotic prescriptions in such a clinical scenario, why would clinicians label the patients as having RTI? For the third recommendation, the pre-existing comorbidities listed are not generic for the susceptibility of infectious conditions. Evidence for pneumonia after acute cough/ acute bronchitis as shown by the fourth recommendation was derived from a prognostic study conducted in the Netherlands among older LRTI patients (≥ 65 years) with the endpoint being 30 day hospitalization or death (Bont et al., 2007), in a region where antibiotic prescription rates were much lower than in the UK. This potentially suggested that more severe cases were included, and the validity of the study results may not apply to a young adult cohort. Also, the study was not primarily designed to predict the onset of infectious complications, for which secondary care is not always needed. Therefore, its clinical implications for UK adult RTI patients in terms of developing complications may not be pertinent.

Table 1.1: National Institute for Health and Care Excellence (NICE) guideline recommendations for immediate antibiotic treatment or investigation for RTI patients (NICE, 2008a)

Offer immediate antibiotics or further investigation/management for patients, who:

are systemically very unwell

have symptoms and signs suggestive of serious illness and/or complications (particularly pneumonia, mastoiditis, peritonsillar abscess, peritonsillar cellulitis, intraorbital or intracranial complications)

are at high risk of serious complications because of pre-existing comorbidity. This includes patients with significant heart, lung, renal, liver or neuromuscular disease, immunosuppression, cystic fibrosis, and young children who were born prematurely.

are older than 65 years with acute cough and two or more of the following, or older than 80 years with acute cough and one or more of the following:

- hospitalisation in previous year
 - type 1 or type 2 diabetes
 - history of congestive heart failure
 - current use of oral glucocorticoids
-

1.6 Rationale for the thesis

1.6.1 Predicting incident pneumonia subsequent to RTIs within 30 days in primary care settings

Pneumonia is one of the most frequent severe infectious complications subsequent to RTIs (Gulliford et al., 2016). In primary care settings, a substantial proportion of pneumonia cases are identified from patients presenting with RTI symptoms (Lieberman et al., 2003). Therefore, pneumonia subsequent to RTIs could be due to delayed diagnosis, untimely treatment, suboptimal medical management or unresolved underlying medical conditions. The onset of pneumonia after incident RTI could be multifactorial, such as diagnosis uncertainty mentioned above; recent changes in host immunity due to wide spread use of immunosuppressant drugs and an aging population; prolonged duration between onset of pneumonia symptoms and clinical pneumonia diagnosis which has been observed among non-severe pneumonia patients under immunosuppressant treatment, heavy alcohol consumption and a priori antibiotic treatment (Sanz et al., 2014b).

In the light of these clinical challenges, it is important to identify risk factors, especially modifiable risk factors, to allow early clinical investigation or prompt treatment to be initiated to decrease the risk of a new episode of pneumonia and facilitate timely preventive interventions. Such preventive procedures will reduce the risk of worsening pre-existing medical condition like COPD or triggering unidentified ill health like cardiac arrhythmias. Meanwhile clinical investigation tests issued based on identified risk factors could offer additional information on undiagnosed conditions such as pulmonary thromboembolism (PTE) or hypersensitivity pneumonitis that would cloud pneumonia diagnosis. To date, risk factors for RTI patients managed in primary care setting who then reconsulted with pneumonia have not been well quantified using primary care data. Existing research based on hospital data does not bear enough validity to outline the pattern in primary care settings. Studies using primary care data alone, or in combination with secondary care data with main measurement being hospitalization due to pneumonia, cannot provide sufficient information to distinguish risk factors affecting patient

hospitalization from those contributing to the onset of pneumonia (Millett et al., 2015).

Atypical pneumonia with early onset being extrapulmonary symptoms is not included in this prognostic study as the diagnosis of this group of pneumonias depend heavily on clinical tests routinely available in hospital settings which go beyond the scope of primary care management. Further, the repeated onset of pneumonia within a predefined time period after an RTI suggests diseases like lung cancer, which is beyond the scope of the main research question for this study.

A time interval of 30 days was chosen to define incident case based on the natural course of common RTIs (NICE, 2008a, NICE, 2019b) after considering the average waiting time for a GP appointment in the UK, and the ease of conventional understanding of infectious complications in the respiratory field.

1.6.2 Inclusive approach for RTI conditions

Both upper and lower RTI conditions together with cough will be included in the adult participant selection since they share similar patient's chief complaints, clinical presentations, clinical investigations and clinical treatment principles in primary care (NICE, 2015, NICE, 2019b, NICE, 2019c). For example, cough has been the leading indication for antibiotic prescriptions among RTI patients (Hawker et al., 2014, Metlay et al., 1998, Little et al., 2017), which could be closely associated with, but not confined to, a single RTI condition. This inclusive selection is intended to reflect daily primary care practice. Additionally, diagnosis drift between subgroups of RTIs has been reported in epidemiological studies in primary care settings (Ashworth et al., 2006, Stocks and Fahey, 2002); therefore, an inclusive selection of target medical conditions will be used to capture the majority of the patient cohort and thus minimize selection bias in the very early stages of the study. Included RTI conditions of the thesis are illustrated in Figure 1.1.

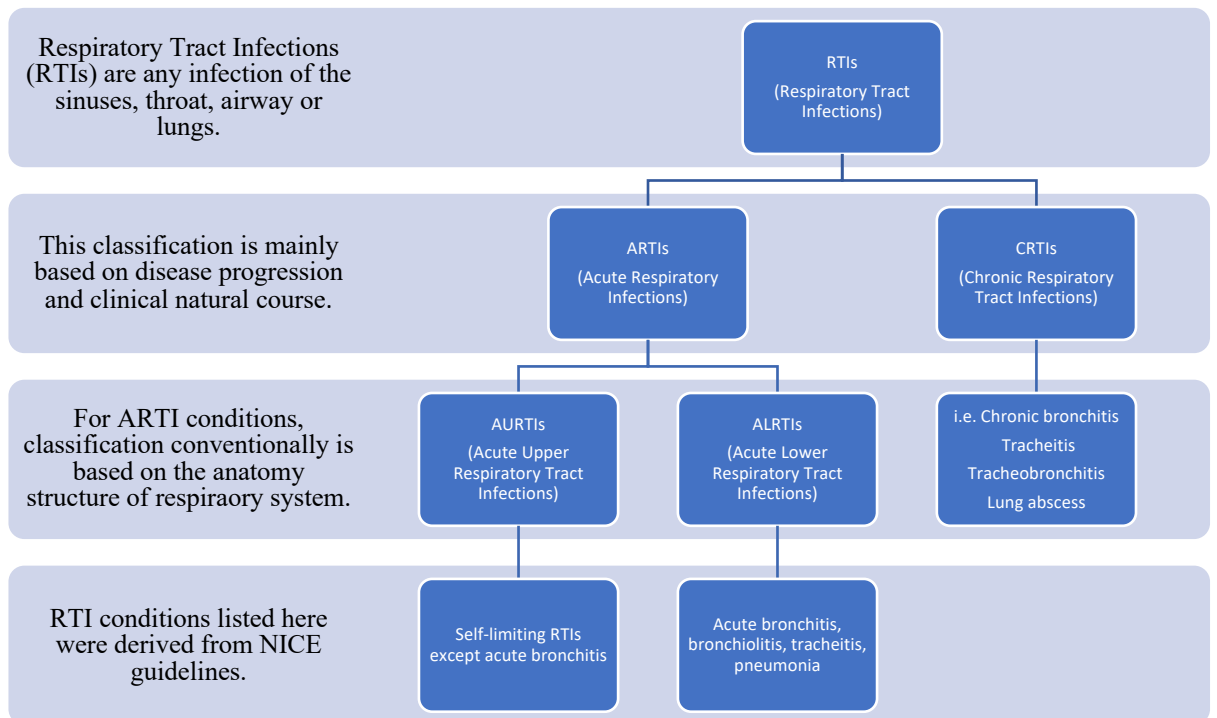


Figure 1.1: Illustration of major RTIs included in this thesis

1.6.3 Study population being adult population

Pneumonia disproportionally affects the young and the elderly; this is primarily determined by the interactions between the host and disease-causal pathogens (Davies et al., 2013, Troeger et al., 2017). In the early stages of life, a child's immune system is not fully developed making them susceptible to many infectious agents (Lewis and Swirsky, 2001, Bradley et al., 2011, Jain et al., 2015b). During adulthood, behavioural factors exert greater influence on the onset and development of infectious diseases; then in the later stages of life, chronic conditions and medical treatments tend to contribute more to the overall outcomes of infection management (Davies et al., 2013, Vinogradova et al., 2009). Also, during daily general medical practice, the clinical features, research evidence and management plans for RTI conditions vary between child and adult groups (Bradley et al., 2011, Jain et al., 2015b, Choi et al., 2013). Research evidence using primary care data for adult pneumonia cases, especially subsequent to RTI among young adult patients in the community settings is limited. Thus, it is hoped that development of a prediction model which follows a life course approach and focuses on the adult age group will provide answers to the research questions for this thesis.

1.6.4 Context of the thesis

This thesis was conducted in the routine clinical primary care settings where most RTI patients were managed in the community settings. The endpoint of prediction modelling was set at the onset on CAP within predefined time period (30 days) irrespective of the severity of the condition or the intensity of treatment. CAP in thesis referred to clinical diagnosed pneumonia in community settings as confirmative chest X-ray are not routinely performed in UK primary care settings. Most predictors were sorted based on clinical guidelines of NICE which provide clinical management reference for medical professionals in the UK. When detailed medical information for certain candidate predictors are not well documented, for example when diabetic patients were identified but the control status of blood sugar was not available or largely missing among patient cohort, data processing for this type of predictors stopped at categorization to gauge the interpretability of study results at the expense of preciseness.

Chapter Two : Thesis research question, aim, objectives and scope

2.1 Research question

What are the characteristics of adult RTI patients managed in primary care settings who reconsulted with pneumonia in 30 days following consultation?

2.2 Research aim

To develop a clinical prediction model for adult RTI patients presenting in the UK primary care settings who reconsulted with pneumonia within 30 days.

2.3 Research objectives

The specific objectives of the thesis are to:

1. Describe and analyse overall antibiotic prescriptions using patient level electronic health records (EHRs) from a subset of participant English general practices from Clinical Practice Research Datalink (CPRD) from 2014 to 2017. Research results served to update existing evidence of antibiotic prescription in the English primary care system, meanwhile provide comparable information to English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) firstly reported in 2014. Antibiotic prescriptions for common infectious conditions managed in primary care including respiratory infections, genito-urinary infections (GUTI), infectious skin conditions, eye infections were evaluated to provide a wide context for the thesis and for data inspection purposes. Also, antibiotics as the major treatment for respiratory infectious conditions managed in the community were investigated from disease management point of view.

2. Conduct a systematic review of previous prediction modelling studies to identify candidate risk factors for pneumonia in community settings. Risk factors identified served to inform variable selection during prognostic model development, and existing prediction modelling methodologies for included studies contributed to inform statistical modelling approaches of the thesis.
3. Conduct an epidemiology study of pneumonia managed in primary care settings in the UK using CPRD data from 2002 to 2017 to map the trajectory of pneumonia incidence and related respiratory conditions. Baseline information on pneumonia and related respiratory conditions in the community contributed to the prognostic study design and interpretation of study results.
4. Develop and validate a clinical prediction model for adult RTI patients who reconsulted with pneumonia within 30 days after their initial consultations. Data source for model development employed CPRD data from 2002 to 2017. Supervised machine learning approaches were adopted to explore possible predictive variables documented in CPRD for the final model. Classification and regression tree (CART), random forest as well as penalized regression methods were examined. Model discrimination performance was assessed using area under receiver operating curve (AUROC), and model calibration capability was examined by Hosmer–Lemeshow test (H-L test) as well. Cross validation was used for internal validation of the final model; temporal validation was performed on the most recent 25% subset of CPRD data.

2.4 Research scope

This research was confined to an adult RTI patient cohort that was managed in primary care settings in the UK. Initial RTI consultations that were not managed during daily routine community care, i.e. out of hours services, private practices, walk in centres and emergency healthcare settings were not explored in this study. Incident pneumonia within predefined time period (30 days) was adopted as the endpoint for prediction modelling; neither repeated pneumonia events after RTI consultation nor pneumonia severity were included in this study. Given that full

external validation and clinical impact studies of prediction model should be delivered by different investigators, prognostic research for this thesis stopped at temporal validation.

This PhD thesis contributed to a wider project on ‘Safety of reducing antibiotic prescribing in primary care: New evidence from electronic health records’ (NIHR award ID: 16/116/46) with CPRD GOLD being the only research data source.

Therefore, electronic health records (EHRs) from CPRD GOLD were deployed as main study data for the whole thesis.

2.5 Structure of the thesis

The thesis is structured in the form of several inter-linked chapters. In chapter three, we define the term ‘electronic health records’ and discuss the strengths and limitations of electronic health records for healthcare research. In chapter four, antibiotic prescription in the English primary care was evaluated, which served to understand respiratory conditions managed in the community from disease treatment point of view. In chapter five, the secular trend in the incidence of CAP together with its three related respiratory conditions were investigated. Chapter six reported a systematic review which aimed to have a scope view of possible candidate prognostic factors for incident CAP to inform model development. Chapter seven provided contextual information of prediction modelling methodology deployed in this thesis followed by chapter eight which detailed and discussed prediction model development as well as model performances. Finally, an overall discussion of study findings of the whole thesis together with reflections gaining through this PhD project were reported.

Chapter Three : Understanding electronic health records (EHRs) and clinical practice research datalink (CPRD)

This chapter presents the overview of EHRs, including their roles in medical research and in primary care research in particular. This is followed by key information concerning the study data set together with details of coding of respiratory conditions for this thesis. The recoding process and results partly contributed to one original research publication (Liang et al., 2018).

3.1 Overview of electronic health records (EHRs)

In medical literature, there is no universally agreed definition of electronic health records (EHRs) with the meaning of the concept varying over time (Häyrinen et al., 2008). The technical definition of EHRs was developed by the International Organization for Standardization (ISO) (Schloeffel, 2004). The key concept is that an EHR is a type of digital data repository with patient healthcare information being the main content, which can be directly accessed by authorized users including healthcare professionals. Retrospective, concurrent and prospective healthcare data can be documented in EHRs and the primary objective for any type of EHR is to document and support healthcare delivery by facilitating quality, efficiency and continuity. Given that EHRs include patient information, data privacy and security considerations should be taken into consideration during construction. Information archived in EHRs are often in the form of unstructured free text, coded data based on standardized terminologies, numerical measurements and test results, or medical images. EHRs have been widely adopted across many healthcare systems during health service delivery but there is often a lack of consistency and interoperability between health systems and levels of care. EHRs may be broadly classified into primary care, secondary care and tertiary care based on the type of healthcare providers with documented information largely reflects the characteristics of corresponding healthcare system, clinical setting, patient cohort, intended users.

3.2 EHRs in medical research

With the advent of 'big data' era, an increasing number of medical research studies have been implemented using EHRs with the primary objective of translating clinical data into better medical knowledge to improve healthcare quality. Big data denotes the large scale of information in terms of volume (scale of sheer volume), velocity (speed of incoming data being generated), variety (different forms of data and data sources), variability (changing nature of these data) and veracity (uncertainty in terms of the accuracy and trustworthiness of the data) (Beam and Kohane, 2018, Rehman and Batool, 2015). EHRs may be considered as big data, not just because of the massive amount of patient data available for analysis but also the broad dimensions of healthcare information being stored (Ross et al., 2014). The amount and complexity of medical information in EHRs sometimes may exceed human capabilities in terms of data collection, processing, and depth of understanding. Novel statistical and data science methodologies, like machine learning, that rely on computationally intensive techniques have been adopted for data mining, manipulation, analysis, interpretation and pattern recognition (Waghlikar et al., 2012, Payne et al., 2010). Given that the primary aim of EHRs is to serve healthcare delivery, documentation of medical information alongside with the provision of healthcare service could be regarded as the main data collection process from research point of view. This enables the secondary use of EHR data for medical research, which may be naturally integrated into EHR systems. However, researchers need to maintain awareness that any factor that potentially affects data recording could exert a consequential influence on data completeness, accuracy, complexity or bias often to an important extent (Hripcsak and Albers, 2012).

In contrast to conventional research approaches where data collection is performed after the research hypothesis or questions have been specified and data definitions agreed, EHRs represent either pre-existing or ongoing collections of medical data produced during routine care. Thus, EHRs are also referred to as 'real life' or 'real world' data and are considered to provide a description of the patients phenotype (Jensen et al., 2012, Sullivan and Goldmann, 2011).

EHRs differ from research data in several important respects. First, data collection for many prospective medical studies is restricted to the period of the research. It is unlikely that this will provide sufficient longitudinal information to map the complex interaction and dynamic nature of human diseases. EHRs offer the potential that multiple measurements of the same trait could be recorded for single individual patients over sufficient timescales to capture chronological trends and better reflect the characteristics of evolving conditions. For example, a recent genome-wide analysis found that the variance in both systolic and diastolic blood pressure that could be explained by single nucleotide polymorphisms (SNPs) was doubled through use multiple measurements of blood pressure from EHRs (Hoffmann et al., 2017).

Secondly, even though population-based registries exist for certain diseases, enabling data collection to be consistent over time, it is not feasible to develop data collection schemes that provide comprehensive coverage of all conditions for patient cohorts or geographic regions, but this is achievable through EHRs (Glicksberg et al., 2018). EHRs might be considered to have better representativeness of actual daily clinical care than disease-specific registries and primary research studies that may have specific inclusion and exclusion criteria. But caution is needed when interpreting the results of studies generated from EHRs, taking into account the real-life circumstances of data collection and avoiding drawing conclusions out of context.

Thirdly, in contrast to purposely collected data, EHRs have the potential to enable discovery of unanticipated associations between variables that do not obviously seem to be related. It is difficult for researchers to investigate underlying associations if relevant information was not included in data collection schemes during trial design for example. For rare conditions, EHRs also offer the possibility to allow precise estimation given quantity of meaningful information contained in the data set. However, it is equally important to be aware of potential false positive findings or analytical results that lack reproducibility generated from analysis of such big datasets (Ioannidis, 2005, Begley and Ioannidis, 2015).

3.2.1 General considerations for EHRs in medical research

Given that EHRs primarily support for care delivery, these basically include ‘stream of consciousness writing’ or a ‘diary’ of healthcare service delivery. From the first contact with patients, to the recording information into EHRs, factors involved in the whole process have the potential to affect and be reflected in the data. Therefore, EHRs are not necessarily direct reflections of patient’s physiological status but are influenced by the complexity of health care process (Tao et al., 2011, Hripcsak and Albers, 2012). This makes it challenging to use EHRs to address medical questions directly. Important issues relating to EHR data research quality include accuracy, completeness, complexity and bias.

Accuracy refers to errors which could be introduced anywhere in the process from first-hand medical information gathering to extracting key information from the raw data, to depositing the information into dataset. Some errors are systematic which introduce bias, as when a better reimbursed diagnosis tends to be more frequently recorded, whereas some errors are random in nature. There are also problems concerning the use of medical concepts, particularly in structured coded data. Even where EHRs function in much the same way as traditional paper medical records that are written manually by individual clinicians, they may not always offer similar consistency between the retrospective interpretation of medical terms and the true intention of the health professional completing the record. This is not because of the recall bias that all retrospective studies may be subject to, irrespective of which type of medical records were used, but standardized clinical terminology systems and ontologies cannot always guarantee that all clinical concepts are coded in the same way as might be expressed in free text information. For example, a question mark in medical notes usually indicates a possible diagnosis based on a doctor’s impressions about patient’s presentations and known medical history. Such uncertainty may not be properly coded, and possible diagnosis could be interpreted as confirmed diagnosis during research, if no additional information is recorded to indicate the level of uncertainty. In addition, there are misalignments between coded concepts and applicable definitions and obvious misuse may occur. For example, 2% of patients with one eye missing were labelled as being ‘equal pupils’ by commonly

used acronym PERRLA in EHRs having ‘pupils equal, round, and reactive to light and accommodation’ (Hripcsak and Albers, 2012). Such apparent misuse of medical term may not always indicate sub-standard health care service: opticians may only perform clinical examination for one eye and default to presuming the other eye’s function is the same, but the interface of structured EHR may not always allow this minority condition to fit into any coded terms.

The **Completeness** of EHR data is generally referred as missingness, but this may not be equivalent to the missing data discussed by statisticians. The main categories of statistical missingness are: missing completely at random (MCAR), there are no systematic differences between missing values and observed values i.e. a measurement was not recorded simply because of computer breakdown; missing at random (MAR), differences between missing and observed values could be explained by observed values i.e. missing blood pressure is expected to be lower than that of observed individuals because younger people may be more likely to miss out blood pressure measurements; and missing not at random (MNAR), systematic differences persist between missing and observed values which could not be explained by observed data i.e. hypertension patients miss out clinical appointments because of exacerbation of the condition like severe headache (Sterne et al., 2009b, Altman and Bland, 2007). Several approaches have been proposed to deal with missing data such as complete case analysis, missing indicator, single and multiple imputation analysis (Groenwold et al., 2012a). Generally multiple imputation might be recommended for data even when MNAR as this reduces bias compared to other methods (Kontopantelis et al., 2017).

For EHRs, data are missing largely not at random for several reasons. Since EHRs mainly contain information relating the episodes of care, information that goes beyond the capture of dataset or data linkage are unlikely to be documented. For example, if patients move between clinical settings during health-seeking process where EHRs are not linked with each other, then EHRs are fragmented due to insufficient clinical information exchange. Also, incident medical episodes mainly depend on patients actively seeking healthcare rather than being followed up. There might be initial onset of certain mild symptoms but recording EHRs is only triggered

when patients present to health care settings. If patients do not attend for medical consultations, then any relevant medical information would not be explicitly recorded before the healthcare episode. In addition, it is possible that items were recorded in error or information was not considered important enough to be documented but might have been of research value.

Complexity is in the nature of healthcare. This makes EHRs data highly complex, including comprehensive individual health information from physical, mental and social well-being aspects, as well as tests and imaging results, and details of the organisational aspects of care. Great efforts are being made to create standardized, hierarchical structured clinical terminology to define clinical data. Such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) which has been deployed in English primary care systems since April 2018 (NHS Digital, 2019b). SNOMED includes more than 300,000 active medical concepts archived under 19 top-level hierarchies (He et al., 2015, Andrews et al., 2007).

Adopting standardized clinical terminology facilitates fast yet large scale healthcare research, however, regional, chronological and temporal variations have been noticed in the use of same clinical terminology system (Pryor and Hripcsak, 1993, Hripcsak et al., 1998, Hripcsak et al., 2005). Apart from coded records, substantial amounts of narrative information, that may contain meaningful information, are stored in the free text of unstructured EHRs. Data documented in the free text may include information on patient symptoms, clinical discretion process, sensitive topic like mental health issues and patient privacy. These narrative records might offer more value for research than coded data, especially for specific research questions but these require natural language processing techniques to extract essential elements (Sevenster et al., 2015, Perotte et al., 2015). Complexity also arises when data are combined from different forms, different settings, across different regions or health care systems (Uddin and Gupta, 2014, Alexander and Wang, 2017).

The concept of **Bias** is important to any type of research data, and especially applies to EHRs. Research studies using EHRs naively, without evaluation of data biases and their magnitudes, may generate research findings that contradict rudimentary and

'common-sense' understanding of healthcare outcomes. For example, a study using EHRs to replicate a prediction study for CAP found higher mortality rates among apparently healthier patient than that in sicker participants (Fine et al., 1997, Hripcsak et al., 2011). Because CAP patients with rapid progression of disease severity died quickly, fewer symptoms were recorded in EHRs which made them appear to be healthier than was truly the case. Another large study evaluating healthcare process impact on the predictive value of EHRs showed that the time at which clinical laboratory tests were ordered predicted survival rates better than the actual laboratory results (Agniel et al., 2018). Such research results are less intuitive, but understandable after factoring in healthcare process: doctors are more likely to order clinical tests for less well or more unstable patients during night shifts, which explain the study results that patients with normal white blood cell (WBC) counts at 4am have lower survival chance than those with abnormal WBC counts at 4pm. Therefore, the nature of EHRs as a composite outcome of various set of healthcare processes introduces the possibility of bias (Boustani et al., 2010).

EHRs contain rich information for medical research but the challenges mentioned above require researchers using EHRs data to have good understanding of the study topic, routine healthcare delivery in that field and to perform thorough data inspection before performing analysis. If uncertainty, inaccuracy, incompleteness and bias of the data are detected, the researcher must determine to what extent this may possibly affect the validity of the study estimates and how study results should be interpreted while recognizing their limitations to ensure the rigour in research. Other practical considerations such as the cost to access a national EHR resource like CPRD should be factored in when conceiving or designing healthcare studies. Because such cost is generally associated with study sample size requested as well as the type of patient information deployed.

3.2.2 Potential usefulness of EHRs in medical research

For the potential usefulness of EHRs, as well as other type of big data in medical research, generally there are four dimensions for consideration. First, EHRs could be used for sketching the outline of disease patterns in population-based study design.

This might lead to two-step approach: using EHRs for variable selection or phenotyping adopting a heuristic, iterative approach to translate raw EHRs data into clinically relevant curated rules; then datasets with more detailed and accurate information with mapping or linkage could be matched to these preliminary rules to facilitate precision medicine. Biomedical studies based on biobanks coupled with EHRs have generated promising results for arrhythmogenic right ventricular cardiomyopathy (ARVC) patients which could be easily diagnosed through electrocardiogram but without notable symptoms (Haggerty et al., 2017, Van Driest et al., 2016). Also, these types of data resources bear the potential to shorten the traditional lengthy drug discovery process by offering novel drug development or repurposing targets (Yao et al., 2011, Shameer et al., 2017, Shameer et al., 2018, Johnson et al., 2017). For example, metformin was suggested to be a potential chemotherapeutic drug contributing to cancer mortality reduction by analysing large EHR data from Vanderbilt University Medical Centre and Mayo Clinic (Xu et al., 2014).

Secondly, data mining of the enriched data resources could allow the identification of underlying medical questions without a research hypothesis being a pre-condition. For example, unsupervised machine learning techniques are able to cluster the data into groups that provide indicative information to inform relevant research questions and service planning. A research study using EHR records following data-mining approach has shown that proton pump inhibitors (PPIs) may be associated with increased risk of cardiac events among the general population (Shah et al., 2015). This type of widely prescribed drug may not necessarily be a causative factor for heart attack but the finding in itself has shown the feasibility and promising utility of analyses in unveiling non-obvious biomedical associations which could be validated by hypothesis-testing studies. Another recent study conducted among type 2 diabetes (T2DM) patients using data from CPRD and the Index of Multiple Deprivation (IMD) data sets has generated informative results through agglomerative hierarchical clustering analysis, even if the IMD data only reflects deprivation information of patient's residential area at LSOA (Lower Layer Super Output Area) level (Nowakowska et al., 2019, CPRD, 2021). The Index of Multiple Deprivation (IMD) combines information from seven domains, including income, employment,

education, health, crime, housing and living environment, to produce an overall deprivation measure. The study identified the healthcare needs for multi-comorbid T2DM patients and highlighted the increasing trends of mental health burden among study population especially the most deprived cohort (Nowakowska et al., 2019).

Thirdly, in addition to outlined disease pattern, specific interventions could be applied to certain modifiable factors of specific healthcare service delivery process and evaluate the corresponding effects promptly. Recently, several studies have provided evidence that computer information algorithms developed through automated techniques like machine learning using quality big EHRs data may outperform medical professionals and be leveraged to improve care (Honeyford et al., 2019, Rajkomar et al., 2018).

Fourthly, information documented in EHRs open the possibilities of certain studies for pragmatic purposes. For example, strong associations after thorough evaluations of confounding variable effects using ample data with comprehensive coverage of possible covariates could function as surrogates of causal relationships especially when novel ground-breaking findings to back up causation are unlikely to be discovered in the near future but expedite decisions have to be done based on current available evidence. One of the most classic example of this kind is the Master Settlement Agreement with the tobacco industry in 1998 (Schroeder, 2004), which has adopted extensive research information to address the strong associations between smoking and its related illnesses to support jurisdiction.

3.3 EHRs in UK primary care research

Properly implemented and well tested electronic health record systems constitute an essential component of electronic health information exchange (HIE) initiative which benefits healthcare delivery from various aspects, further, offering great potential for medical research. Efficient information exchange between healthcare providers improves the safety of service delivery through rapid oversight of reliable consistently recorded medical histories by medical professionals (Kaelber and Bates, 2007). The accuracy and consistency of population healthcare information with

satisfactory coverage entail the research value of EHR data comparing to segmented or self-reported patient data. In addition, healthcare information exchange backed up by EHRs has potential to empower patients for better joint medical decision making, optimize healthcare resource efficiency as well as continual service improvement through learning health systems (Tang et al., 2006, Payne et al., 2010, Pagliari et al., 2007, Abernethy et al., 2010, Deeny and Steventon, 2015). EHR system could enable studies with patient involvement to be conducted efficiently at large scales.

During the past five decades, one of the major achievements of national-scale health informatics investment in the UK is that almost 100% of the primary care system is now covered by EHRs (Payne et al., 2010), at a time when only a small fraction (10-30%) of ambulatory care clinicians in north American (the US and Canada) were using EHR routinely (Jha et al., 2008). Such computerisation subsidised by government investment has become the prerequisite of a number of healthcare delivery frameworks to enhance primary care system performance and research like the Quality and Outcomes Framework (QOF) (NHS Digital, 2021).

The research value of EHR in UK primary care is endorsed by the universal health coverage system (Rivett, 2014) where 98% of the population is registered with a general practice and GP visits under the National Health Service (NHS) are free of charge (Herrett et al., 2015). These imply the representativeness of the data as well as minimization of healthcare insurance reimbursement impact during healthcare delivery. Another possible benefit is that better interoperability of healthcare information exchange could be achieved under such a healthcare system where major barriers introduced by fragmented services delivery model i.e. in the US due to insurance-based health-care processes could be avoided (Blumenthal and Dixon, 2012). Additionally, the so called 'from cradle to grave' healthcare service in the UK, where a unique NHS number (its equivalent being the Community Health Index (CHI) Number in Scotland) is assigned to individual patients, offered both clinical and research value (Boyd et al., 2018, NHS Research, 2020). Such unique patient identifier enables consistent documentation of distinct patient healthcare information within the UK's health and social care system and data linkage establishment, which is of particular research value for longitudinal population study in the community.

Information documented in primary care EHRs shares different characteristics from secondary healthcare data. Hospital EHRs cover accurate and granular information of episodic albeit intense clinical treatment, whereas patient information maintained in primary care EHRs comprise comprehensive patient information ranging from social, demographic and lifestyle information to disease management records. Despite the debate around the quality of primary care data in comparison with hospital EHRs as well as other conventional medical data i.e. disease specific registry and laboratory data, the longitudinal life span health care information contained within these EHR systems in themselves demonstrate the research value that may seldom be rivalled by other medical data of any kind (Johnson et al., 2014). Recent research evidence has demonstrated that primary care medical records could provide complementary information to hospital care data. A recent prognostic study conducted among intensive care units (ICUs) patients in Denmark using EHR data showed that a machine learning model generated from aggregated clinical data: including longitudinal disease history before admission from primary care data and real-time physiological measurement from hospital data, outperformed existing mortality prediction rules as well as the ones derived from any individual type of data alone. Even previous clinical diagnoses dated back to 10 years before ICU admission has shown to retain predictive value (Nielsen et al., 2019). Therefore, primary care data here have provided independent information much like stable baseline descriptions of patient health conditions, which assisted to differentiate underlying reasons even if similar acute physiological statements were presented, further, to inform clinical treatment and outcome prediction. An analogy of research values offered by primary and secondary care data in medical research is to contrast studies about regional climate and weather of individual days. Therefore, primary care data and secondary care data could be suitable for many research purposes individually but bridging primary and secondary care data could exploit routinely collected healthcare data to address certain types of specific research questions with enhanced cost-effectiveness.

Apart from top-down initiatives, software and hardware vendors also played an essential role during successful implementation of EHRs in the UK primary care (Johnson et al., 2014). In England, there are mainly three GP EHR systems: Vision,

EMIS and SystmOne (NHS, 2016, Kontopantelis et al., 2018). Each GP IT system possesses individual clinical coding classification systems which should be taken into consideration when multiple datasets were merged for data aggregation. For example, both the Clinical Practice Research Datalink (CPRD) Gold database and the Health Improvement Network (THIN) database collect data from general practice using Vison, whereas general practices contributing data to CPRD Aurum adopted EMIS system (Wolf et al., 2019, Franklin and Thorn, 2019). Awareness should be given to possibilities that variability in coding styles may persist between and within GP IT systems. For example, the documentation of mortality events, a clinical code could be applied in one EHR system, but same information may be stored in a sperate specific field under another system. Further, the deployment of EHR systems has shown to be geographically clustered with appreciable regional variability in the English primary care (Kontopantelis et al., 2018). Therefore, the representativeness of EHR data generated from a single GP IT system should be considered when attempting to generalize study results.

3.4 The clinical practice research datalink (CPRD) Gold

3.4.1 Description of CPRD Gold

The data source for this thesis were obtained from the UK CPRD GOLD dataset which contains anonymised EHRs data from primary care. The database was originally constructed in London in 1987, known as the small Value Added Medical Products (VAMP) dataset and expended into General Practice Database (GPRD) in 1993 (Kousoulis et al., 2015, Williams et al., 2012). In 2012, with the allowance of linkage to secondary care data and incorporation of more mortality together with demographic data, the dataset was referred as clinical practice research datalink (CPRD). Due to its wide geographic coverage (England, Wales, Scotland and Northern Ireland) and ongoing data collection since 1987, CPRD became the largest primary care EHR dataset of longitudinal medical records in the world with 5.1 years follow-up duration on average (Herrett et al., 2015). A subset of participant English practices (approximately 75%, accounting for 58% of all UK CPRD practices) consent to contribute data to the CPRD linkage scheme. Currently available linkages

include Hospital Episode Statistics (HES) (secondary care data) , Office for National Statistics (ONS) (mortality and cause of mortality data), Index of Multiple Deprivation (IMD), Townsend scores and Carstairs Index (deprivation data), Mental Health Dataset (MHDS) (mental health data) and disease registries like the cancer registry, data from Public Health England (PHE). For other available linked data as well as planned linkage, relevant information can be accessed via <https://www.cprd.com/linked-data> (CPRD, 2019a). Recently, a decreasing number of GP practices participating CPRD has been noted due to the declining use of Vision system. In October 2017, CPRD started a new dataset named CPRD Aurum collecting data from consenting GP practice from England and Northern Ireland (data from Northern Ireland is available from 2019) using EMIS system (Wolf et al., 2019). Previous CPRD dataset with primary care data from Vision system is then referred as CPRD Gold (CPRD, 2019b). Given that primary care data sources for this thesis were derived from CPRD Gold or subset of CPRD Gold, the dataset will be referred to as CPRD for the remainder of the thesis.

3.4.2 Main data files of CPRD and data quality

The CPRD is one of the largest primary care databases worldwide with ongoing collection of anonymised medical records from approximately 700 general practices covering more than 11.3 million patients. There were 4.4 million active patients whose data met quality criteria contributing data to the dataset (Herrett et al., 2015). CPRD comprises comprehensive information on patients' demographics, medical conditions, medication prescription, clinical management, clinical investigation tests, referrals and clinical findings as shown in Table 3.1.

Given that 98% of the UK population is registered with a general practice and CPRD includes about 6.9% of the UK population (Herrett et al., 2015), the population coverage of the dataset was broadly considered to be representative in terms of age, sex, ethnicity and geographical distributions (Herrett et al., 2015, Van Staa et al., 2001).

Internal data quality measurements were carried out at both patient and practice level. Such internal data quality assessment efforts aimed to ensure data quality meet satisfactory level within its subsets of UK practices (Williams et al., 2012). Individual patient data were checked against a list of key variables indicating data validity, so that patients' data labelled as 'acceptable' were recommended for research purpose. Practice level assessment was manifested by the 'up-to-standard' (UTS) status based on two essential concepts: consistent data documentation and mortality rates within acceptable range (Williams et al., 2012, Jick et al., 1992). It has been noted that more rigorous data quality assessment rules are warranted to characterise the strength and weakness of primary care data for research, since influential regulatory incentives like Quality Outcome Framework (QOF) rules have changed the data recording practice which may not be captured by UTS parameters (Vamos et al., 2011, Tate et al., 2011).

The validity of research evidence from CPRD has been extensively studied especially for diagnosis and drug prescription. The majority of these studies examined data validity with reference to external research data sources such as GP questionnaire, Office for National Statistics (ONS), GP data documentation requirements (Harshfield et al., 2017, Williams et al., 2012). A systematic review including 212 publications from CPRD showed high estimates of validity with a median of 89% cases being confirmed (Herrett et al., 2010). Similarly, less than 1% difference in smoking prevalence was reported between CPRD and Health Survey for England (Booth et al., 2013). However, minor discrepancies of death recording between CPRD and ONS data were reported (Harshfield et al., 2017, Delmestri and Prieto-Alhambra, 2020) and misclassification in body mass index (BMI) (Bhaskaran et al., 2013) have been identified suggesting variance in data validity may occur when specific research questions are intended to be addressed using CPRD data. The quality of CPRD data must be reconsidered from the perspective of each new study.

Table 3.1: Main data files of CPRD (CPRD, 2017)

File	Content
Patient file	Basic patient demographics and patient registration details for the patients
Practice file	Details of each practice, including region and collection information
Staff file	Practice staff details with one record per member of staff
Consultation file	Information relating to the type of consultation as entered by the GP from a pre-determined list. Consultations can be linked to the events that occur as part of the consultation via the consultation identifier (consid)
Clinical file	Medical history events. This file contains all the medical history data entered on the GP system, including symptoms, signs and diagnoses. This can be used to identify any clinical diagnoses, and deaths. Patients may have more than one row of data. The data is coded using Read codes, which allow linkage of codes to the medical terms provided.
Additional Clinical file	Information entered in the structured data areas in the GP's software. Patients may have more than one row of data. Data in this file is linked to events in the clinical file through the additional details identifier (adid).
Referral file	Referral details recorded on the GP system. These files contain information involving patient referrals to external care centres (normally to secondary care locations such as hospitals for inpatient or outpatient care), and include speciality and referral type
Immunisation file	Details of immunisation records on the GP system
Test file	Records of test data on the GP system. The data is coded using a Read code, chosen by the GP, which will generally identify the type of test used. The test name is identified via the Entity Type, a numerical code, which is determined by the test result item chosen by the GP at source. There are three types of test records, involving 4, 7 or 8 data fields (data1 - data8). The data must be managed according to which sort of test record it is. Data can denote either qualitative text-based results (for example 'Normal' or 'Abnormal') or quantitative results involving a numeric value.
Therapy file	Details of all prescriptions on the GP system. This file contains data relating to all prescriptions (for drugs and appliances) issued by the GP. Patients may have more than one row of data. Drug products and appliances are recorded by the GP using the Gemscript product code system.

3.5 Recoding of RTI, chest infection, pneumonia and pleural infection

Clinical terminology is an essential element for EHR research especially when coded data are employed as the main study resource. Understanding the context of code selection and how clinical information were documented under structured classification systems relates to case definition and further data extraction based on refined code lists (Gulliford et al., 2009b). Code lists are the starting point for many epidemiological studies using EHRs, as without knowing the baseline for certain medical conditions, any further analysis would be built up without a solid foundation. Code list identification for this thesis started off with comparisons between clinical terminology/ disease classification systems followed by finalization of RTI, pneumonia and plural infection code lists.

3.5.1. Comparisons between major diseases classification systems

The comparisons being made in this section aimed for formulizing a general framework to tailor code lists to thesis research objectives and providing contextual information for critical appraisal EHR study results from various clinical terminologies in similar filed. Read code and International Classification of Primary Care (ICPC) were adopted to illustrate primary care terminologies whereas ICD-10 was adopted to illustrate that of secondary care. Other disease classification systems for primary or ambulatory care such as International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) adopted in insurance-based healthcare systems like the US (CDC, 2020) were not included in this thesis. Comparison results are presented in ((A refers to Appendix) Table A 1, Table A 2 and Table A 3).

GPs in the UK recorded medical events to structured EHRs formerly using the Oxford Medical Information System (OXMIS) (Gulliford et al., 2009b), latterly the Read code system (Williams et al., 2012), then moved onto SNOMED CT since April 2018. During the transition time before SNOMED CT is mandated by 1st April 2020, Read code remains to be held in clinical IT systems together with SNOMED

CT as dual coding which allows existing clinical and research to continue (NHS Digital, 2019c). Read code version 2 (5-byte READ) was one of the standard clinical terminologies being used in the UK primary care settings during study period with mapping to SNOMED CT and ICD-10 (NHS Digital, 2019a, NHS Digital, 2015). Read code follows a semi-hierarchical structure with more than 100,000 codes to document routine clinical consultations. These codes were mapped onto ‘medical codes’ as documented in CPRD for the identification of medical events, diagnoses, referral events and certain treatment issued. Clinical section of 5-byte version was a hierarchy list of 5-digit codes with 0-9 and A-Z (except I, O, V and W) in the first place and 0-9 in the remain four places and plus up to four trailing period ‘.’ characters. The relative position of individual code to another is indicated by itself: [H....] is the common ancestor of respiratory system diseases with ‘H’ being the first character, and [H0...] in turn the common ancestor of all acute respiratory infections begins with ‘H0’.

The International Classification of Primary Care Second edition (ICPC-2) is another classification system (WHO, 2020) applied to general practice in countries like Australia (Faculty of Health Sciences, 2018). It has 17 chapters, 15 out of which represent the localization of the complains and/or diseases for body systems with the remaining two chapters being unspecified conditions and social problems. This allows health care providers to use with ease by allocating the episodes to relevant categories. Each chapter is structured by seven components which mainly deal with consultation reasons for encounter, clinical investigation and findings, interventions or procedures in the disease management process. The taxonomy is based on three-digit alphanumeric codes with the first letter representing the organ or body system defined by those 17 chapters followed by two digits of numerical codes representing information relating to the seven components for each chapter. Such coding system would allow the capture of episodes of care (EoC) over time and consistent documentation of clinical information by coding all the consultation episodes relate to the same condition (WONCA, 2016). In this case, ICPC does not have a separate section for general symptoms which have already incorporated in each episode of disease management. Whereas, for read codes the medical history, clinical symptoms, disease diagnosis and clinical investigations are separately documented in

the system which would allow the analysis of various aspects in the medical industry of public or clinical importance in contrast with that of ICPC.

ICD is a standard clinical diagnostic terminology developed by WHO designed to serve the needs for hospital care, health management and epidemiological purpose (WHO, 2019a). For now, ICD-10 5th edition comprising 22 chapters has been mandated in the UK since 2016 onwards (HSCIC, 2015). ICD primarily aims to serve hospital health care where patients commonly present for a single episode of visiting with a chief problem clearly differentiated in most scenarios. This classification largely reflects clinician's perspective on individual's illness which may not necessarily continue to apply to primary care where patient's chief complain may not be specific and clinicians are managing multiple episodes of the same patients over time. Thus, ICD with such a detailed granularity limits its clinical usefulness in primary care.

On the other hand, Read code and ICPC would allow many health problems and even non-disease conditions to be labelled at a sufficient detail meanwhile the overarching terms in the primary care coding systems enable primary care data aggregation feasible. Also, these coding systems are not solely diagnosis oriented but taking elements from patient's part which is essential for patient centred care. By employing coding systems suitable for primary care, health care provider can efficiently document daily care practice reflecting the content of primary care. This further assist health care providers, medical researchers and policy makers to understand what is happening within primary care system and how to improve services.

However, this does not mean ICD stands a position in competition with primary care coding systems, rather they are complementary to each other, for example expanded codes through data mapping between ICD and primary care terminologies would separate out certain conditions included in a high-level aggregated code into more specific codes for research purposes.

3.5.2 Recoding RTI, chest infection, pneumonia and pleural infection

The identification of RTI consultations relates to the selection of both participants and cases, therefore inclusive recruitment of all the possible codes with multi-layered

recoding criteria would allow the sensitivity analysis and coding drift assessment to be conducted. The code list sorting started with an existing list. Complementary information was then sought after from CPRD dictionary. Two researchers (MG and XS) reviewed and agreed on the final list. Respiratory tract infections were firstly coded into 10 categories as shown in Figure 3.1: acute otitis media (72 codes), acute sore throat/acute pharyngitis/acute tonsillitis (103 codes), common cold (9 codes), influenza (15 codes), acute rhinosinusitis (22 codes), acute cough/acute bronchitis (74 codes), URTI (4 codes), chest infection (6 codes), LRTI (19 codes), pneumonia (103 codes), plus pleural infection (50 codes). Pleural infection codes were used for approximate estimation of severe pneumonia cases (Sun et al., 2019). Apart from that, the rest 10 subdivisions of recoded conditions were further grouped into self-limiting RTIs (299 codes), chest infection (25 codes), pneumonia (103 codes), pleural infection (50 codes). Finally, 427 codes comprising all RTI conditions were identified in consistent with previous study using the same data set (Ashworth et al., 2004a).

For pneumonia code list sorting, natural language process techniques were also explored to validate human code extraction (Liang et al., 2018). Satisfactory agreement as measured by F statistics being 0.84 between human search and algorithm generated code lists was reported. This has provided recoding implications for future studies using EHR coded data, especially when well established code lists are not readily available for reference and potential medical concepts have been documented sparsely in clinical terminology systems.

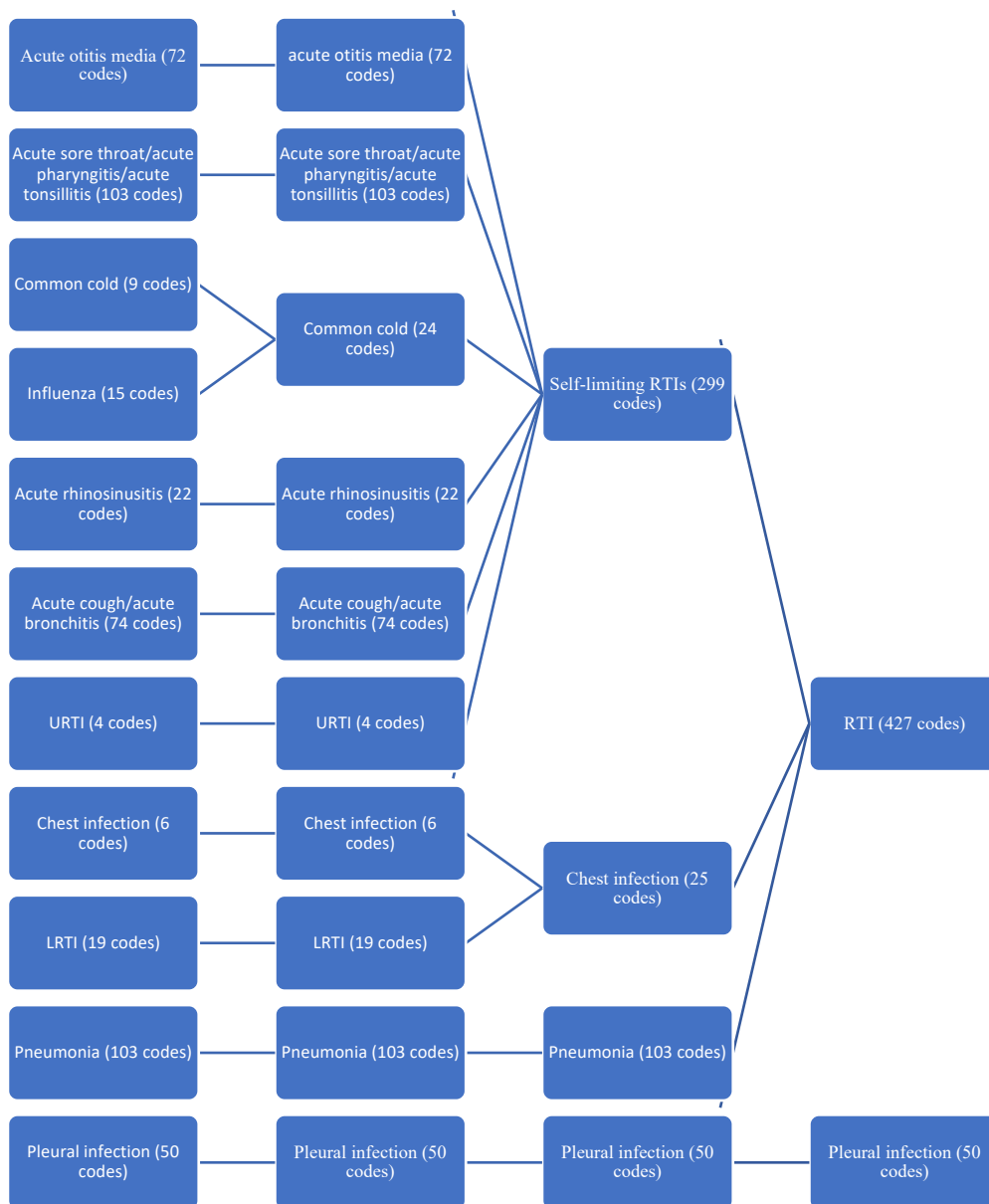


Figure 3.1: RTIs and pleural infections recoding process in CPRD

Chapter Four : Reducing antibiotic prescribing in primary care in England from 2014 to 2017: a population-based cohort study

This chapter presents an epidemiological study to investigate the changes in antibiotic prescribing in primary care in England. An original research paper was published from this study (Sun and Gulliford, 2019).

4.1 Introduction

Antibiotics have paved the way for numerous modern medical achievements by effectively preventing and treating infection conditions, which would be life threatening before the first introduction of penicillin in late 1940s (Gaynes, 2017). This immediate and profound initial impact introduced the ‘golden era’ of antibiotic discovery between 1930s and 1960s (Nathan and Cars, 2014). ‘Golden’ denotes that such success appeared to be usual at that time and ‘era’ means that such a time has now faded away. Apart from the great challenges and major investments, the therapeutic potential of current antibiotics is now shrinking because of the ever-increasing emergence of antimicrobial resistance (AMR), which becomes the major disincentive. Almost all the disease-causing pathogens have developed resistance mechanisms to antimicrobial agents which are used for treating them (Ruiz et al., 2012).

The reasons underpinning the high rates of antibiotic resistance at population level are multifactorial but AMR is a direct consequence of extensive medical utilisation (Hellen Gelband, 2015, Laxminarayan et al., 2013), especially in the community settings (Control and Prevention, 2013, Costelloe et al., 2010, Goossens et al., 2005). Due to the severe challenges caused by antibiotic resistance, infectious diseases have come to under the spotlight again as a major public health problem in modern medicine. With the advent of carbapenem-resistant Enterobacteriaceae (CRE), modern medicine would likely proceed to post-antibiotic era with a depleted anti-infectious arsenal (WHO, 2011, Reardon, 2014). Therefore, curbing AMR has

become a public health priority. Global and national antibiotic stewardship campaigns have been launched to improve judicious use of antibiotics and contain the spreading of antibiotic resistant microbials (WHO, 2013, Davies and Gibbens, 2013, Stone et al., 2012, Schwaber et al., 2011, Brinsley et al., 2005, MacDougall and Polk, 2005, So and Woodhouse, 2010).

With the increasing understanding of negative effects of antibiotic misuse in the short term and long run, national policy initiatives in the UK have been launched to lower unnecessary use of antibiotics (Ashiru-Oredope et al., 2012b). Multifaced interventions together with public campaigns have be promoted to reshape the antibiotic prescribing patterns since 1998 (Carbon and Bax, 1998, Rodgers et al., 1999, Ashworth et al., 2002).

A continuous decreasing trend for antibiotic consumption has been reported since 1995. By the year of 2000, significant reduction of antibiotic prescription had been achieved with the driving force being decreased antibiotic utilization in the community settings (Ashworth et al., 2004a) where 70% to 80% of antibiotics were dispensed (PHE, 2016, PHE, 2015, PHE, 2014, Committee). The English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) (PHE, 2017a) reported that antibiotic prescriptions in primary care have decreased by 13% between 2012 and 2016.

The purpose of the present study is to update data for antibiotic prescribing trends in English general practices from 2014 to 2017. Antibiotic prescriptions for males and females and for patients of different ages were evaluated. Also, antibiotic prescriptions for common categories of indications were analysed to provide estimates for detailed antibiotic utilizations in the community. Finally, trends in prescribing of broad-spectrum β -lactam antibiotics and individual classes of antibiotics were compared.

4.2 Methodology

4.2.1 Data source and main measures

For this study, a subset of CPRD data from 102 general practices in England which participated in the data linkage scheme, and consistently contributed data in all years from 2014 to 2017 were analysed. This serves to eliminate the impact of practices contributing data in different years. During study period, the total number of general practices in the UK contributing to CPRD declined from 491 in 2014 to 285 in 2017. Meanwhile, the number of CPRD general practices in England participating the data linkage scheme declined from 257 to 102 (Table 4.1). Individual participant data were included from the later of 1st January 2014 or the start of the patient's CPRD record to the earlier of 31st December 2017 or the end of the patient's CPRD record. Data were obtained from the February 2018 release of CPRD. For practices that ended CPRD data collection during 2017, an equivalent end-of-year-date was also adopted for earlier years, because of the marked seasonality in antibiotic prescriptions.

For each year of study, person-time was employed as denominator for rates. Person-time contributed by each individual patient was calculated between 1st January of the year, or start of registration if this was later, to 31st December of the year, or end of registration or date of death, if these were earlier. This is to ensure that the study deployed active patient records rather than historical records.

Prescriptions for antibiotics were identified using product codes for all antibiotic drug classes included in section 5.1 of the British National Formulary (BNF) except anti-tuberculous, anti-lepromatous agents and methenamine, which were excluded (BMA, 2013, PHE, 2017a). The BNF groups antibiotic drugs into the following categories: penicillins, cephalosporins (including carbapenems), tetracyclines, aminoglycosides, macrolides, clindamycin, sulphonamides (including combinations with trimethoprim), metronidazole and tinidazole, quinolones, drugs for urinary tract infection (nitrofurantoin) and other antibiotic drugs. There is no universally accepted definition for 'broad-spectrum' antibiotics (BMA, 2013). For this study, broad-

spectrum β -lactam antibiotics included broad-spectrum penicillins, cephalosporins. Carbapenems are only rarely used in primary care and were combined with cephalosporins for further analysis. Clinical indications for antibiotic prescription were grouped into categories based on Read medical codes recorded into patients' clinical and referral records on the same date as the antibiotic prescription including: 'respiratory conditions', 'genitourinary conditions', 'skin' conditions, 'eye' conditions, non-specific codes recorded, no codes recorded. Codes employed for analysis are shown in Appendix C. An inclusive approach was adopted to code selection in order to capture any potential indications for antibiotic treatment. The 'issueseq' field in the CPRD therapy file was used to evaluate whether prescriptions were repeat prescriptions. Prescriptions associated with 'issueseq' values of zero were coded as 'acute' while 'issueseq' values of one or above were coded as 'repeat' prescriptions.

Table 4.1: Numbers of general practices contributing to CPRD from 2014 to 2017. Figures are mid-year counts.

Variable	2014	2015	2016	2017
All CPRD general practices	491	422	338	285
CPRD general practices in England	329	260	180	133
CPRD general practices in England participating in data linkage	257	202	238	102

4.2.2 Statistical analysis

Prescriptions for all antibiotic and broad-spectrum β -lactam antibiotic were enumerated by year. Antibiotic prescriptions of the same type on the same date were considered as a single event. Age was included as a continuous variable but was also analysed in sub-groups from 0-4 years, then 10-year age-groups up to 85 years and over. Read codes recorded on the same date as an antibiotic prescription were analysed according to indication. The primary indication on each date was allocated by giving priority to indications in the following sequence: respiratory, genitourinary, skin and eye. Antibiotic prescription rates were calculated as per 1,000 person years, and proportions of registered patients with antibiotic prescribed in year in relation to age-group, gender, study year and main indications. In order to estimate annual changes in antibiotic prescribing, we fitted in hierarchical generalized linear Poisson models using the 'hglm' package (Rönnegård et al., 2010) in the R program. The dependent variable was a count of antibiotic prescriptions (either all antibiotic prescriptions or broad-spectrum antibiotic prescriptions). Predictors were calendar year, sex and age, including quadratic and cubic terms to allow for non-linear effects of age. Calendar year was included as a linear predictor based on inspection of descriptive data and because non-linear effects would be difficult to estimate over a four-year period. A random effect for general practice was included because of the repeated observations on general practices over years. The log of person-time was included as offset. Relative rate reductions were estimated as one minus the adjusted relative rate for the linear effect of calendar year. In view of the size of the dataset, we present confidence intervals rather than significance tests. Results were presented using the 'ggplot2' and 'forestplot' packages (Wickham, 2016a) in the R program (Team, 2013).

4.2.3 Data governance approval

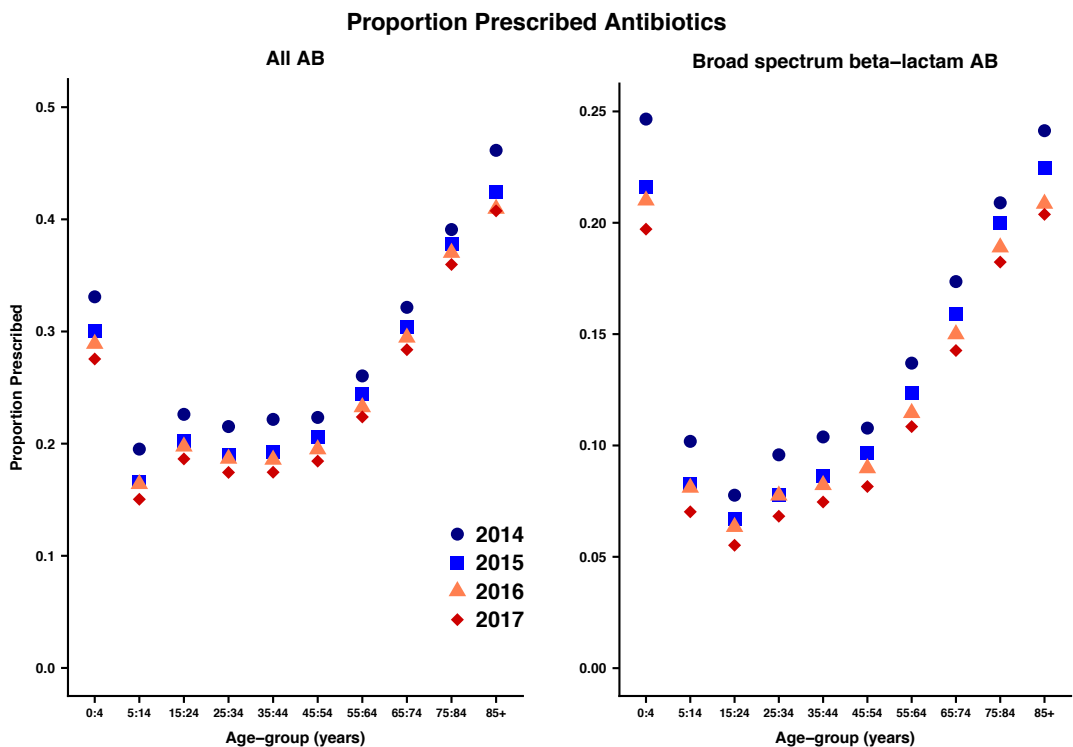
Research protocol for this study was submitted to and approved by the Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC), which is responsible for reviewing all proposals in

CPRD. Research Approval was obtained on 10th February 2016 (Protocol NO. 16_020). All patients' EHRs analysed for this study were completely anonymized.

4.3 Results

4.3.1 Overall antibiotic prescriptions

Analyses included 102 general practices that contributed data in each year from 2014 to 2017 as shown in Table 4.2. The registered population was 1.03 million in 2014 increasing to 1.07 in 2017. There were 539,219 antibiotic prescriptions in 2014, declining to 459,476 in 2017. The antibiotic prescribing rate declined from 608 per 1,000 in 2014 to 489 per 1,000 in 2017. The proportion of registered patients that were prescribed antibiotics in year declined from just over 1 in 4 (25.3%) in 2014 to just over 1 in 5 (21.1%) in 2017. Figure 4.1 (left panel) shows changes in the proportion of patients prescribed antibiotics in year over the study period. A consistent year-on-year reduction was observed in each age-group from 0 to 4 years to 85 years and over. Marked antibiotic prescribing variations were observed in relation to age, with the highest rates at the extremes of age.



Note: Given that the confidence intervals of presented proportions are too narrow and hence less informative, confidence intervals are not illustrated in this figure.

AB: antibiotic

Figure 4.1: Proportion of patients prescribed antibiotics in year by age-group and calendar year.

There were 195,750 broad-spectrum β -lactam AB prescriptions in 2014, declining to 153,423 in 2017. The proportion of all AB prescriptions that were broad-spectrum β -lactams decreased from 36.3% in 2014 to 33.3% in 2017 (Table 4.2). Figure 4.1 (right panel) shows the change in proportion of patients prescribed broad spectrum β -lactam AB by age-group. While there was a year-on-year decrease in broad-spectrum β -lactam antibiotic use in each age-group, the absolute reduction appeared to be greater at older ages in whom broad-spectrum β -lactam AB use was greatest.

Table 4.3 presents data for antibiotic prescribing indications. Respiratory consultations accounted for the most frequent defined indication with 168,852 (31%) prescriptions in 2014 and 129,032 (28%) in 2017. The most frequent respiratory codes for antibiotic prescription were ‘cough’ and ‘chest infection’ as shown in

Table A 4. Genitourinary infections and skin infections accounted for 9% and 7% of antibiotic prescriptions respectively with little change over years. There were 77,431 (14%) antibiotic prescriptions with no associated medical codes recorded in 2014 and 73,596 (16%) in 2017. There were 204,395 (39%) of antibiotic prescriptions with only non-specific codes recorded in 2014 and 181,018 (39%) in 2017. Frequently used non-specific codes ‘telephone encounter’, ‘patient reviewed’, ‘had a chat to patient’ and ‘administration’.

Table 4.3 shows the proportion of repeat prescriptions for different prescribing indications. In 2017, 78,166 (17%) of antibiotic prescriptions were recorded as repeat prescriptions. The proportion of repeat prescriptions was 2% or lower for respiratory, genitourinary or eye conditions. For skin infections, 8% of antibiotic prescriptions were recorded as repeat prescriptions. There were 10% of repeat prescriptions among antibiotic prescribing episodes associated with non-specific codes. Among 73,596 antibiotic prescriptions in 2017 with no medical codes recorded, 56,216 (76%) were recorded as repeat prescriptions.

Table 4.2: Numbers of antibiotic prescriptions, and antibiotic prescribing rates, by year. Figures are frequencies except where indicated.

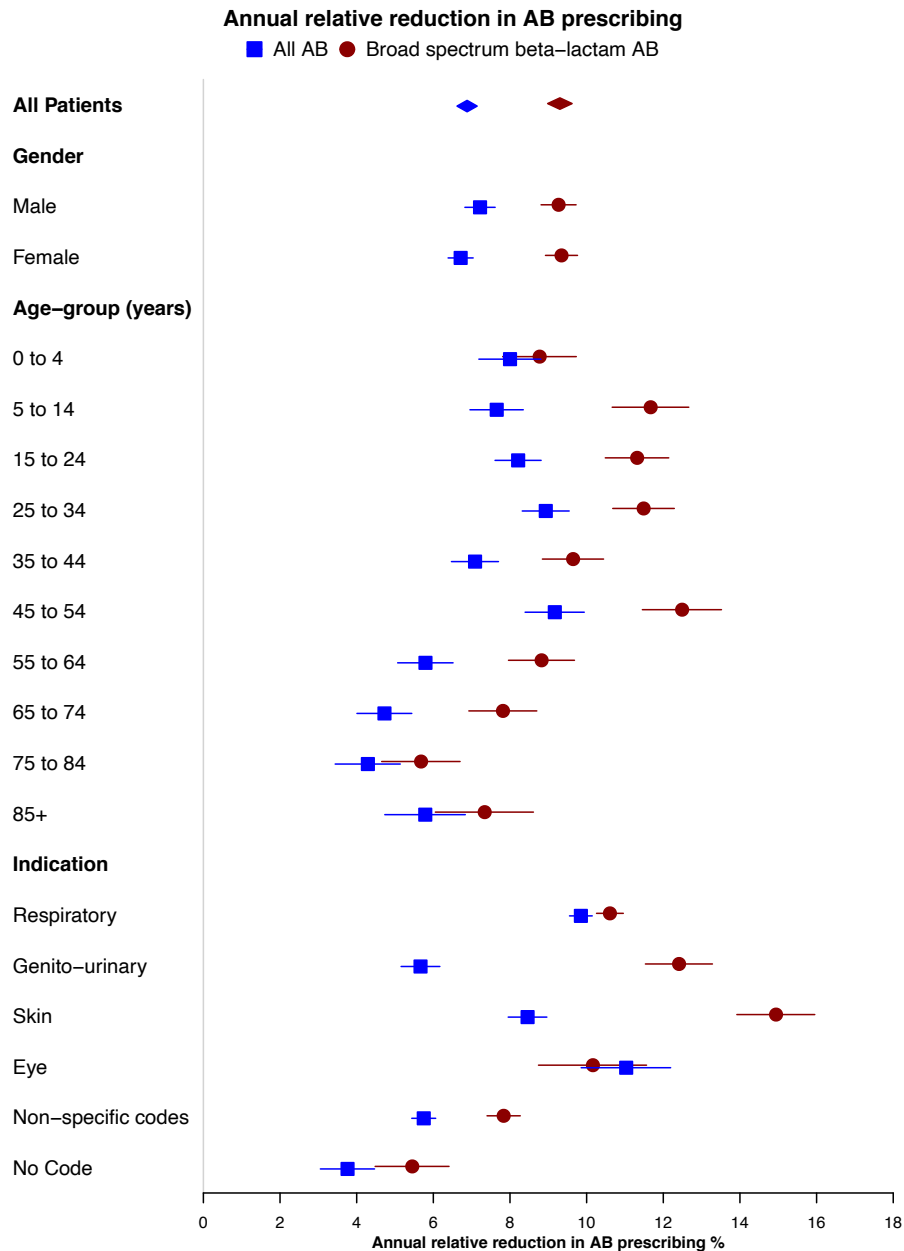
	2014	2015	2016	2017
General practices	102	102	102	102
Patients	1,025,539	1,058,805	1,069,513	1,071,293
Female (%)	520,336 (50.7)	536,082 (50.6)	542,051 (50.7)	543,324 (50.7)
Age (mean, sd, years)	39.4 (23.4)	39.5 (23.4)	39.7 (23.5)	39.9 (23.5)
Person-time (person years)	887,580	921,735	932,544	939,620
All antibiotic prescriptions	539,219	494,185	482,917	459,476
All AB prescribing rate (per 1,000 person years)	608	536	518	489
Proportion of patients prescribed AB (%)	25.3	23.0	22.2	21.1
Mean number of AB prescriptions in patients prescribed	2.08	2.03	2.03	2.03
Broad-spectrum β-lactam antibiotic prescriptions	195,750	174,353	167,056	153,423
Broad-spectrum prescribing rate (per 1,000 person years)	221	189	179	163
Proportion of patients prescribed Broad-spectrum β -lactam AB (%)	12.9	11.3	10.7	9.9
Mean number of BS-AB prescriptions in patients prescribed	1.48	1.46	1.45	1.45

Table 4.3: Distribution of antibiotic prescriptions by broad groups of indications. Figures are frequencies except where indicated.

	2014		2015		2016		2017		Acute ^a	Repeat ^a
	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq. (%)	Freq. (%)
AB prescriptions	539,219		494,185		482,917		459,476		381,310 (83)	78,166 (17)
Respiratory conditions	168,852	31.3	146,025	29.5	140,263	29.0	129,032	28.1	127,474 (99)	1,558 (1)
Genito-urinary conditions	47,009	8.7	44,544	9.0	42,453	8.8	42,401	9.2	41,740 (98)	661 (2)
Skin conditions	39,579	7.3	35,299	7.1	33,640	7.0	32,003	7.0	29,513 (92)	2,490 (8)
Eye conditions	1,953	0.4	1,622	0.3	1,586	0.3	1,426	0.3	1,399 (98)	27 (2)
Non-specific codes	204,395	38.0	191,565	38.8	189,386	39.2	181,018	39.4	163,804 (90)	17,214 (10)
No medical codes	77,431	14.3	75,130	15.2	75,589	15.7	73,596	16.0	17,380 (24)	56,216 (76)

a: Proportion of antibiotic prescriptions that were either acute or repeat prescriptions in 2017

Informed by the apparent consistent annual declines in antibiotic prescribing noted in Table 4.2 and Figure 4.1, Figure 4.2 presents a Forest plot of annual relative reductions in antibiotic prescribing adjusted for age, gender and general practice. Estimates for all antibiotic prescribing are shown in blue and for broad-spectrum antibiotic prescribing in red. The annual relative reduction in all antibiotic prescribing was 6.9% (95% confidence interval 6.6% to 7.1%). Estimates were generally similar for males and females. For participants aged less than 55 years, the sub-group estimates were all greater than the overall estimate, being greatest at age 45 to 54 years at 9.2% (8.4% to 9.9%) per year. For participants older than 55 years, estimates were consistently lower than the overall estimate being lowest at age 75 to 84 years and above at 4.3% (3.4% to 5.1%) per year. Considering sub-groups of indications, rates of decline were greatest for respiratory indications (9.8%, 9.6% to 10.1%), and eye indications (11.0%, 9.9% to 12.2%). The rate of decline was smallest for antibiotic prescriptions with no recorded indication (3.8%, 3.1% to 4.5%). The overall rate of decline was faster for broad-spectrum antibiotics than all antibiotics at 9.3% (9.0% to 9.6%). Estimates were consistent for males and females. The greatest relative decline was observed at 45 to 54 years (12.5%, 11.5% to 13.5%) and the lowest at 75 to 84 years (5.7%, 4.7% to 6.7%). The greatest decline was for skin condition indications (14.9%, 13.9% to 15.9%) and lowest for un-coded indications (5.5%, 4.5% to 6.4%).



Note: AB: antibiotic

Figure 4.2: Forest plot showing annual relative reduction (95% confidence interval) in antibiotic prescribing for all antibiotics and broad-spectrum β -lactam antibiotics between 2014 and 2017 for sub-groups of age and gender and different prescribing indications. Estimate were adjusted for age, gender and clustering by practice.

4.3.2 Changes in different classes of antibiotics

Figure 4.3 presents changes over time in the utilisation of different classes of antibiotics. The most frequently issued antibiotics were penicillins, accounting for 56% of antibiotic prescriptions in men and 44% in women in 2017; macrolides, men 14%, women 12%; tetracyclines, men 14%, women 12%; sulphonamide and trimethoprim combination, men 6%, women 11%. the latter class be more frequently used in females. Clindamycin, aminoglycosides and other antibiotics accounted for less than 1% of antibiotic prescriptions and are not shown. During the period of study, drugs for urinary tract infections (nitrofurantoin) increased as a proportion of all antibiotic prescriptions, in men from 2.6% in 2014 to 4.2% in 2017, and in women from 8.8% in 2014 to 13.7% in 2017. Tetracycline use also increased between 2014 and 2017, in men from 12.8% to 14.5% and in women from 10.1% to 11.6%. Most other categories appeared to show slight declines. Both penicillin and macrolides were mainly prescribed for treating respiratory conditions, whereas tetracyclines was frequently issued for skin conditions among young patients and respiratory conditions in later life. There was a decline in the use of sulphonamide/trimethoprim combinations for urinary conditions while a notable increase of nitrofurantoin use for these conditions was observed over study years among all age groups but more particularly in women.

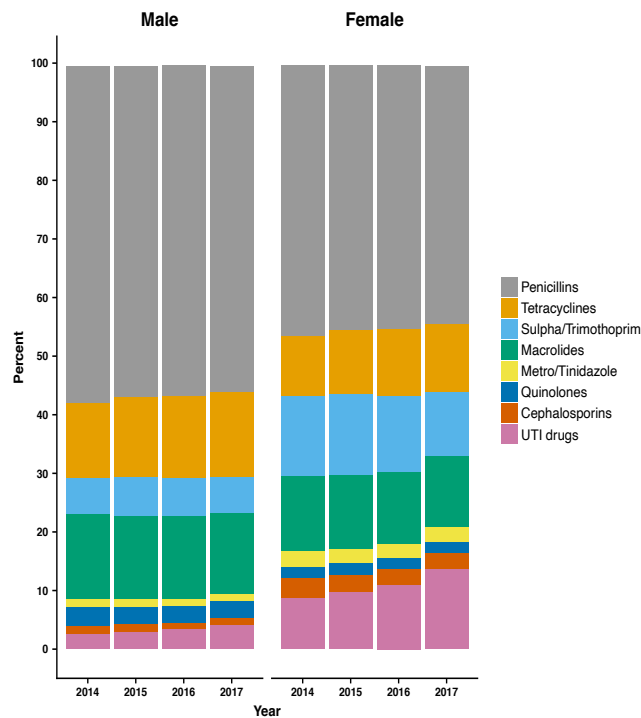


Figure 4.3: Bar chart showing changes from 2014 to 2017 in the proportion of antibiotic prescriptions for different antibiotic classes for males and females.

4.4 Discussion

4.4.1 Main findings

The rate of antibiotic prescriptions and the proportion of patients receiving antibiotics have declined consistently over this four-year period. Antibiotic utilisation shows important patterning by age and gender, being higher in very young and very old people and higher in women than men. However, the results show that a reduction in antibiotic utilisation is being achieved across all ages groups and in females as well as males. The gender gap in relation to antibiotic prescribing could be due to differences in medical care-seeking behaviour or specific conditions which disproportionately affect one gender (Smith et al., 2018).

Among prescriptions associated with coded indications, respiratory conditions were the most frequent indication for antibiotic prescription and also showed the greatest rate of decline. Respiratory tract infections have always been the ‘bread and butter’ among common conditions managed by GPs in the UK, and antibiotics are issued at about 50% of RTI consultations overall (Gulliford et al., 2014b). In general, paralleled decreasing trends have been observed between RTI consultation rates and antibiotic prescription rates during the past two decades (Ashworth et al., 2004a, Frischer et al., 2001). Selective antibiotic prescription decreases among subgroups of RTI were also detected with greater reduction took place among URTIs, thus may offset the insignificant decline among LRTIs in general (Ashworth et al., 2004a, Ashworth et al., 2005, Gulliford et al., 2009a). Given the wider social and medical context, GPs tend to withhold antibiotic prescriptions to low risk patients, and consequently URTI consultation rates have been declining in primary care (Ashworth et al., 2005). It becomes essential that judicious use of antibiotics takes two major aspects into consideration: reducing unnecessary use of antimicrobials to reserve the anti-infection treating potentials, while timely administrating anti-infectious treatment when necessary to optimize the therapeutic effects and, further avoiding undesirable infectious complications. Therefore, clinical safety outcomes should be analysed alongside with the reduction outcomes of antibiotic prescription (McDonagh et al., 2016, Gulliford et al., 2016).

Consistent with other recent reports (Dolk et al., 2018), a substantial proportion of antibiotic prescriptions were found no association with specific coded clinical indications. Antibiotic prescriptions that were not associated with medical codes, showed the slowest rate of decline, potentially further identifying this category of prescriptions as representing a sub-optimal standard of clinical practice which might hamper the accurate estimation of drug indications. This might also be due to the possibility that relevant medical information like clinical symptoms were documented in free texting data, which are no longer accessible in CPRD since 2013 (Wolf, 2018). However, these results suggested that prescriptions without coded indications included a high proportion of repeat prescriptions. Enhancing the quality

of clinical information recording is warranted in order to improve patient care, as well as the usefulness of records for research and health service management.

More than one third of prescriptions were for β -lactam antibiotics and there was evidence of an important decline in antibiotic prescribing in this category consistent with previous evidence (PHE, 2017a). The relative reductions of broad-spectrum β -lactam prescriptions were greater than for overall antibiotic utilisation. Broad-spectrum β -lactam antibiotics may not necessarily offer more effective coverage of causal pathogens than their more specific counterparts. The present results suggest that clinicians are gradually shifting to more targeted narrow-spectrum substitutions when possible. There is no universally agreed definition for ‘broad-spectrum’ antibiotics, therefore this study employed the category of β -lactam antibiotics that were broad-spectrum (as ‘broad-spectrum beta-lactam antibiotics’) to illustrate the possible difference in prescribing trends between these broad-spectrum antibiotics and their counterparts. For most common and uncomplicated infections, narrower spectrum drugs are generally recommended as first-line agents in general practices (PHE, 2017b). Macrolides are generally recommended as substitutions for penicillin in the case of penicillin allergy, as well as for specific indications including *Legionella* or the eradication of *Helicobacter pylori* (HP). Nevertheless, macrolides were frequently prescribed in this and other studies (Aabenhuis et al., 2017, van den Broek d'Obrenan et al., 2014). Clinical use of tetracyclines was low in children in recognition of the risk of deposition in growing bone and teeth but the overall use of tetracyclines was higher at other ages (Association, 2013). The increase of nitrofurantoin utilization was mainly due to the shift of guideline recommendation from trimethoprim to nitrofurantoin as empiric treatment for genitourinary conditions (PHE, 2017b).

4.4.2 Strengths and limitations

The study included more than 100 general practices in England that participated consistently across the four-year period of study. The CPRD includes general practices from throughout the UK. However, because the CPRD licence imposes limits on the size of dataset to be employed, only CPRD general practices in England

were included in this study. During the period of the study, there was substantial attrition of the cohort of CPRD general practices as practices migrated from the Vision practice systems that was employed by practices contributing to the CPRD database. It was considered to be important to include the same general practices in each year of study, with more than 100 general practices included in total. However, it may not be certain whether the antibiotic prescribing of general practices that left the CPRD might differ from those that remained, nor study results could be generalized into all general practices in England.

Previous studies have demonstrated the high quality and completeness of primary care electronic health records in CPRD. The data suggested that repeat antibiotic prescriptions might account for a high proportion of uncoded prescriptions but data in the 'issueseq' field has not been well-validated to our knowledge. A concern for the present study is the possible lack of recording of out-of-hours prescriptions, especially those from deputising services, walk-in centres and emergency care settings (Williams et al., 2017). It is noted that codes for telephone consultations and home visits were frequent among antibiotic prescriptions with non-specific coded indications, which suggests that some out-of-hours activity may have been captured.

We also acknowledge that prescriptions from hospitals and specialist clinics are not included, but these are expected to make only a small contribution to community antibiotic utilisation. It appears unlikely that the large and consistent reductions in prescribing observed in this study could be accounted for by shifting of prescribing to other care settings. The research analysed prescriptions issued and not prescriptions dispensed or consumed by patients. We were not able to determine whether prescribers used a delayed or deferred antibiotic prescribing strategy. For these reasons, we believe that actual antibiotic consumption may be slightly lower than we have reported. We acknowledge that there are variations in prescribing between practices (Dolk et al., 2018, Ashworth et al., 2005, Pouwels et al., 2018) our analytical method allowed us to estimate overall effects, and measures of precision, that accommodated variation between practices. The study results show some difference from an earlier study (Dolk et al., 2018) in terms of distribution of

indications, but since different general practices, from different databases, were included in the two studies this may reflect variations in clinical practice.

4.4.3 Comparison with other studies

Previous analyses of primary care electronic health records have focused on antibiotic prescribing for respiratory infections (Ashworth et al., 2004b, Gulliford et al., 2009a) recognising that these conditions represent the most frequent indications for antibiotic prescription. There has been a long-term decline in respiratory consultation rates in England that has contributed to reducing antibiotic utilisation for these conditions (Ashworth et al., 2004b). Some authorities suggest that respiratory consultations account for nearly 60% of antibiotic utilisation in primary care (National Institute for and Clinical, 2008). Our analyses are consistent with those of Dolk et al.(Dolk et al., 2018), who found that respiratory consultations account for fewer than half of antibiotic prescriptions. However, a high proportion of prescriptions may be associated either with no medical codes or non-specific codes making interpretation difficult. There were further methodological differences between the Dolk et al.(Dolk et al., 2018) study and our own, the former study relied on the THIN database with a different number of general practices participating in different years, as well as using code lists that may have differed in some respects. Consequently, minor numerical differences are to be expected.

4.5 Conclusion and implication for future research

The present analyses add to recent reports by providing age- and gender-adjusted estimates of the rate of decline in antibiotic utilisation for all antibiotics and broad-spectrum β -lactam antibiotics, for different prescribing indications and different population sub-groups defined by age and gender. The results show that the recent decline in antibiotic prescription is broadly based and has been observed in all sub-groups investigated. However, the decline in antibiotic utilisation has been at a faster rate for broad-spectrum antibiotics than all antibiotics; the decline is consistent by gender but tended to be lower over age 55 years; the slowest rate of decline is observed for antibiotic prescriptions with no coded indications.

The results emphasise the utility of electronic health records for providing individual-patient data for surveillance of trends in antimicrobial utilisation and focusing future efforts at antimicrobial stewardship where these are most needed. Meanwhile, bias introduced by uninformative codes or clinical information documented elsewhere such as free texting, should be investigated to gauge the rigor of research results.

Chapter Five : Pneumonia incidence trends in UK primary care from 2002-2017: a population-based cohort study

This chapter presents an epidemiological study to investigate the chronological changes in pneumonia incidence and related respiratory conditions in UK community settings. An original research paper was published from this study (Sun et al., 2019).

5.1 Introduction

Community acquired pneumonia (CAP) remains a public health concern globally especially among older and young population (Prina et al., 2015, Welte et al., 2012, Podolsky, 2005, Brown, 2012). As an ambulatory care sensitive condition (ACAS) (Frick et al., 2017), CAP can be effectively managed in primary care or it may sometimes result in hospitalization with the place of treatment largely depends on severity assessment (NICE, 2014, Lim et al., 2009a, Buendia et al., 2010). Recently, increasing hospital admissions due to pneumonia have been reported by several studies across various regions including U.S., UK and Spain (Fry et al., 2005, Quan et al., 2016a, Trotter et al., 2008, Vila-Corcoles et al., 2009, Kaplan et al., 2002). Such increasing trends may not be fully explained by ageing populations nor the increase of comorbid conditions (Trotter et al., 2008, Quan et al., 2016a). In Oxfordshire in the UK, hospital admission for adult CAP increased from 1998 to 2014 with accelerating rate after 2008. Increasing trends were similar across all age groups (Quan et al., 2016a). Other researches on emergency hospital admission for pneumonia also suggested the increasing trends (Blunt, 2013, Bardsley et al., 2013).

Existing evidence has indicated that the burden of pneumonia on hospital care is growing but this may not necessarily suggest that the incidence of CAP is increasing among the general population. Pneumonia patients managed in primary care settings may have different disease profile from those requiring inpatient care because decisions on hospital admission for pneumonia patients mainly depends on clinical evaluation of disease severity, general health status as well as the patient's social circumstances (Lim and Woodhead, 2011, Levy et al., 2010, Metlay et al., 2019).

Additionally, pneumonia itself may trigger exacerbations of underlying health conditions or deteriorate pre-existing ill health, such as chronic heart failure or coronary syndromes (Eurich et al., 2017, Corrales-Medina et al., 2010) to deteriorate and require hospitalization. Therefore, mapping the temporal changes of pneumonia and relevant respiratory conditions managed in the community is warranted to provide baseline information for further studies.

In primary care, respiratory infections management may often be ‘symptom oriented’ as definitive diagnosis confirmed through clinical investigatory tests are less readily available compared with secondary care. CAP cases may often be identified based on clinical features rather than from clinical investigations including radiology findings and bacteriological tests (Lim et al., 2009b). As discussed in chapter there, general practice in the UK adopt the Snomed CT (NHS Digital, 2019c) and Read Code classifications (NHS Digital, 2015), which enable coding of comprehensive and detailed patient information including occupation, social circumstances, clinical symptoms and signs, clinical tests and use of medical services. This is in contrast to the disease categorizations offered by the International Classification of Diseases (ICD) (World Health Organization, 2018) used for coding of hospital episode statistics. Therefore, less specific codes including ‘chest infection’ could be adopted to label clinically-suspected pneumonia during clinical consultations. Clinical guidelines in the UK emphasise that the category of ‘chest infection’ may contain two general clinical scenarios: acute bronchitis, and CAP, with antibiotics only being indicated for the latter (NICE, 2015). Conversely, there may be no major differences between management recommendations for CAP and clinically-suspected pneumonia labelled as ‘chest infection’ in terms of essential elements of treatment, severity assessment and referral principles in adult population (NICE and Excellence, 2014, NICE, 2015). From a disease management perspective, adult CAP and antibiotic treated chest infection cases might be labelled interchangeably during consultations in primary care.

To have a good understanding of pneumonia management, it is necessary to evaluate trends in the incidence of conditions treated as pneumonia in the community. This will provide complementary information to hospital-based research evidence. In this

study, CAP was included as ‘clinically-diagnosed pneumonia’ whereas antibiotic-treated chest infection as ‘clinically-suspected pneumonia’. Influenza pneumonia was included to investigate whether secondary pneumonia complicating influenza has exerted significant impact on pneumonia burden in the community (Metersky et al., 2012, Chan, 2009). Pleural infection was also analysed as a proxy of severe pneumonia because it shares a similar aetiology to pneumonia and may be considered to be a direct infectious complication subsequent to some CAP cases (Brown, 2012, Dean et al., 2016). Increasing trends in pleural infection incidence as well as hospitalization rates were reported (Grijalva et al., 2011) especially among children (Metersky et al., 2012, Krenke et al., 2016).

5.2 Methodology

5.2.1 Study design and data source

This population-based cohort study was conducted using data from the CPRD with detailed information described in chapter three. The present analysis included all eligible family practices contributing to CPRD, and data for all registered patients aged up to 100 years old, during sixteen calendar years from the beginning of 2002 to the end of 2017.

5.2.2 Case definition

Four respiratory infectious conditions were included in the analysis: pneumonia, antibiotic treated chest infection, influenza pneumonia and pleural infection. Read code lists for included conditions were reviewed and selected by two researchers with clinical and epidemiological backgrounds.

Pneumonia was identified using Read codes associated with ‘pneumonia’ term after excluding tuberculosis (TB), fungal, parasite pneumonia and influenza pneumonia. The remaining pneumonia codes were refined further so that non-infectious pneumonia codes like Bronchiolitis obliterans organizing pneumonia (BOOP) (Epler, 2001) were excluded. This was in line with the group of CAP being mainly adopted

to label uncomplicated bacterial pneumonia in primary care. CAP was then evaluated as ‘clinically-diagnosed pneumonia’ in this study. Influenza pneumonia was analysed as a separate category. Antibiotic-treated chest infections were defined when chest infection cases received antibiotic prescriptions on the same date of clinical consultation. Antibiotic prescriptions issued to chest infection patients include antibacterial agents from ‘chapter 5.1 of the British National Formulary without anti-viral, anti-TB, anti-leprosy and anti-fungal drugs (Committee, 2017). Antibiotic treated chest infection was evaluated as ‘clinically-suspected pneumonia’ in this study. Empyema and bacterial pleurisy were grouped into pleural infection during analysis. Codes for the same condition in the same patient within 90-day time-window were considered to represent a single episode.

5.2.3 Statistical analysis

Both crude incidence and standardised incidence rates were calculated during chronological trends evaluation. Incident events were considered as those recorded more than one year after the patient start-date to eliminate prevalent cases from any possible duplication of records during patient registration. Person-time at risk was estimated for the CPRD registered population by year from 2002 to 2017 as denominator for crude incidence. For each individual participant of eligible practices, contributing person-time included from the latest of the patient registration date, or the date the family practice began contributing data to CPRD, to the earliest of the patient end-of-registration, death date or the date the practice left CPRD. Age-specific incidence rates were calculated using the age-groups 0-4 years, 5-14 years, then 10-year age-groups, up to 85 years and older. Data for those aged above 100 years were not included in this study due to limited number of cases in this group as well as possible data recording errors. Age- and sex-standardised incidences (ASR) were calculated using the European standard population (2013 revision).

Recent chronological trends of disease incidence were modelled by joinpoint regression analysis (Kim et al., 2000) using Joinpoint Trend Analysis Software from the NIH National Cancer Institute (Statistical Research and Applications Branch, 2018). The joinpoint method starts with a simple linear model (zero joinpoint) and

tests whether addition of joinpoints improves goodness of fit using Monte Carlo permutation tests with methodological details reported elsewhere (Kim et al., 2000, Kim et al., 2004). The minimum and the maximum number of joinpoints are set in advance, but the final number of joinpoint(s) together with the location of time point(s) are determined statistically. In this study, the minimum number was set to be 0 and the maximum specified to be 3 as recommended by the user manual of Joinpoint Trend Analysis software (NIH, 2018). Annual percent changes (APC) were estimated to quantify the direction and slope of the trend in given period of time between two joinpoints and Average Annual Percent Change (AAPC) was adopted to measure average rate changes across the whole study period. Joinpoint regression analysis has been applied to identify temporal trends especially when the quantity of interest is incidence, prevalence or mortality rates (Doucet et al., 2016, Chaurasia, 2020b, John and Hanke, 2015) with calendar year generally being the time scale (Akinyede and Soyemi, 2016, Chatenoud et al., 2016).

5.2.4 Sensitivity analyses

Analyses were repeated using data from the 218 general practices which contributed data to CPRD consistently across all study years from 2002 to 2017. This aimed to evaluate whether changes in the general practice population exerted significant influence on study conclusions.

5.2.5 Data governance approval

The research protocol for this study was submitted to and approved by the Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC), Protocol 16_020.

5.3 Results

There were 550 general practices contributing to CPRD in 2002, increasing to 631 in 2007, before declining to 314 in 2017 (Table 5.1). A total of 4.2 million person-years of follow up of registered patients aged up to 100 years was identified in 2002, increasing to 5.03 million in 2008-2009, before declining to 2.5 million in 2017.

5.3.1 Clinically-diagnosed pneumonia

The number of episodes of clinically-diagnosed pneumonia was between 5,000 and 10,000 in each year of study (Table 5.1). Six codes accounted for 81% of all pneumonia episodes: 'pneumonia due to unspecified organism' (38%); 'bronchopneumonia due to unspecified organism' (14%); 'history of pneumonia' (12%); 'community acquired pneumonia' (8%); 'lobar (pneumococcal) pneumonia' (7%); 'lobar pneumonia due to unspecified organism' (3%). The crude incidence rate of clinically-diagnosed pneumonia increased from 1.50 (1.46 to 1.54) per 1,000 patient years in 2002 to 1.64 (1.60 to 1.68) per 1,000 in 2010, the rate then increased more rapidly to 2.22 (2.16 to 2.28) per 1,000 in 2017. Figure 5.1 (upper left panel) shows changes in the age-standardised rates of pneumonia for men and women separately with fitted lines from joinpoint regression. Trends were similar in men and women but clinically-diagnosed pneumonia was more frequent in men. Table 5.2 presents estimates from the joinpoint regression model. The annual percent change (APC) in age-standardised rate of clinically-diagnosed pneumonia was 0.3% (95% confidence interval -0.6 to 1.2%) per year from 2002 to 2010 but from 2010 to 2017 the APC was 5.1% (3.4 to 6.9%) per year. The average annual percent change over the entire period was 2.4% (3.4 to 6.2%) per year. Estimation of age-specific rates of pneumonia (Figure 5.2) shows that clinically-diagnosed pneumonia was reducing throughout the period in children aged under 15 years, while recorded clinically-diagnosed pneumonia increased in adults, especially at older ages. In patients aged 15 to 54 years, rates of clinically-diagnosed pneumonia were slightly higher in women than men but, but over the age of 55 years clinically-diagnosed pneumonia was more frequent in men, especially at the oldest ages.

5.3.2 Clinically-suspected pneumonia

The annual number of cases of clinically-suspected pneumonia ranged between 44,662 and 152,992. Two codes accounted for more than 99% of clinically-suspected pneumonia: ‘chest infection not otherwise specified’ (61%) and ‘chest infection’ (39%). The crude rate of clinically-suspected pneumonia was more than 10 times higher than for clinically-diagnosed pneumonia, increasing from 23.7 in 2002 to 30.4 per 1,000 in 2008 before declining to 18.2 per 1,000 in 2017 (Table 5.1). Figure 5.1 (upper right panel) shows that trends age-standardised rates of clinically-suspected pneumonia were similar in men and women, but absolute rates were greater in women than men. Joinpoint regression indicated a change in trend in 2008. The APC from 2002 to 2008 was 3.8% (0.8 to 6.9%) per year compared with -4.9% (-6.7 to -3.1%) per year after 2008. Changes in age-specific rates were generally consistent but clinically-suspected pneumonia was more frequent in women from 15 to 74 years but more frequent in males during childhood and over the age of 75 years. The overall trend for all chest infection diagnoses was similar to clinically-suspected pneumonia with the same turning point of 2008. While the proportions of all chest infections that were clinically-suspected pneumonia increased steadily from 66% in 2002 to 88% in 2017 at an average APC of 1.6% (1.4%-1.8%) per year, suggesting that ‘chest infection’ not treated with antibiotics declined more rapidly than treated chest infection.

5.3.3 Influenza pneumonia and pleural infection

Table 5.1 and Figure 5.1 present data for influenza pneumonia and pleural infection (including bacterial pleurisy and empyema). Rates of influenza pneumonia showed a peak in 2009 but remained low in other years. Rates of pleural infection were low and showed no consistent trend over time. Pleural infection was more frequent in men than women but there was no gender difference for influenza pneumonia.

Table 5.1: Number of incidence events of pneumonia and related conditions. Figures are frequencies except where indicated.

Year	Family practices	Person Years	Clinically- diagnosed pneumonia		Clinically-suspected pneumonia		Influenza pneumonia		Pleural Infection	
			Freq.	Rate ^a	Freq.	Rate ^a	Freq.	Rate ^a	Freq.	Rate ^a
2002	550	4,191,630	6,291	1.50	99,256	23.68	956	0.23	136	0.03
2003	586	4,458,959	7,189	1.61	116,781	26.19	1081	0.24	157	0.04
2004	612	4,767,931	7,219	1.51	117,265	24.59	1264	0.27	157	0.03
2005	620	4,898,961	8,184	1.67	139,785	28.53	1473	0.30	186	0.04
2006	626	4,956,288	7,812	1.58	134,030	27.04	1454	0.29	192	0.04
2007	631	5,016,169	7,798	1.55	151,411	30.18	1468	0.29	200	0.04
2008	627	5,025,191	8,043	1.60	152,922	30.43	1500	0.30	207	0.04
2009	621	5,026,729	7,977	1.59	133,848	26.63	6103	1.21	215	0.04
2010	613	4,967,771	8,145	1.64	136,905	27.56	2110	0.42	210	0.04
2011	596	4,862,957	8,141	1.67	127,963	26.31	1574	0.32	177	0.04
2012	580	4,805,309	8,529	1.77	133,994	27.88	1309	0.27	196	0.04
2013	564	4,595,318	8,376	1.82	112,730	24.53	1025	0.22	177	0.04
2014	530	4,201,387	7,802	1.86	96,219	22.90	859	0.20	134	0.03
2015	462	3,605,006	7,353	2.04	76,108	21.11	672	0.19	107	0.03
2016	371	2,861,175	6,346	2.22	57,126	19.97	536	0.19	102	0.04
2017	314	2,455,307	5,457	2.22	44,662	18.19	430	0.18	91	0.04

^arate per 1,000 person years

Table 5.2: Joinpoint regression estimates for annual percent change (APC).

Condition	Measure	Year of joinpoint	APC (%) before joinpoint (95% CI)	APC (%) after joinpoint (95% CI)	Average APC (%) 2002 to 2017 (95% CI)
Clinically-diagnosed pneumonia	Crude-rate	2010	0.4 (-0.7 to 1.5)	4.8 (3.4 to 6.2)	2.4 (3.4 to 6.2)
	ASR	2011	0.3 (-0.6 to 1.2)	5.1 (3.4 to 6.9)	2.2 (1.4 to 3.0)
Clinically-suspected pneumonia	Crude-rate	2008	3.9 (0.9 to 7.1)	-4.7 (-6.6 to -2.9)	-1.4 (-2.8 to 0.1)
	ASR	2008	3.8 (0.8 to 6.9)	-4.9 (-6.7 to -3.1)	-1.5 (-2.9 to -0.1)

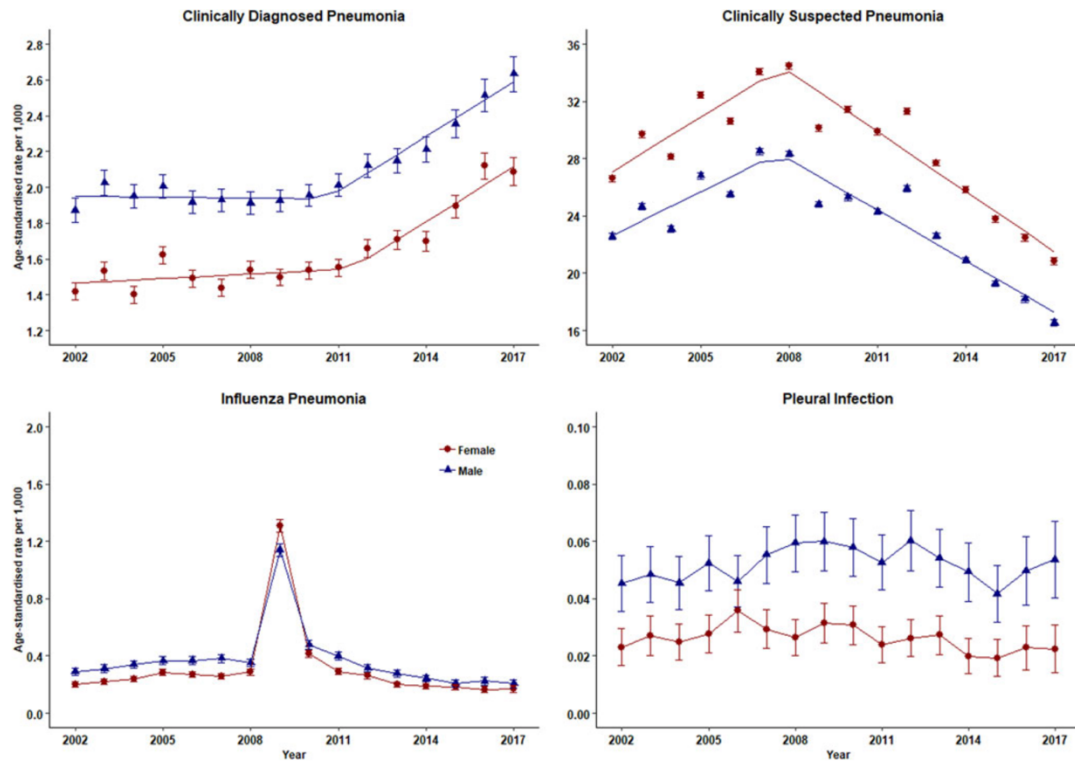


Figure 5.1: Trends in pneumonia and related conditions for both men (blue) and women (red) 2002-2017. Rates are per 1,000 person-years.

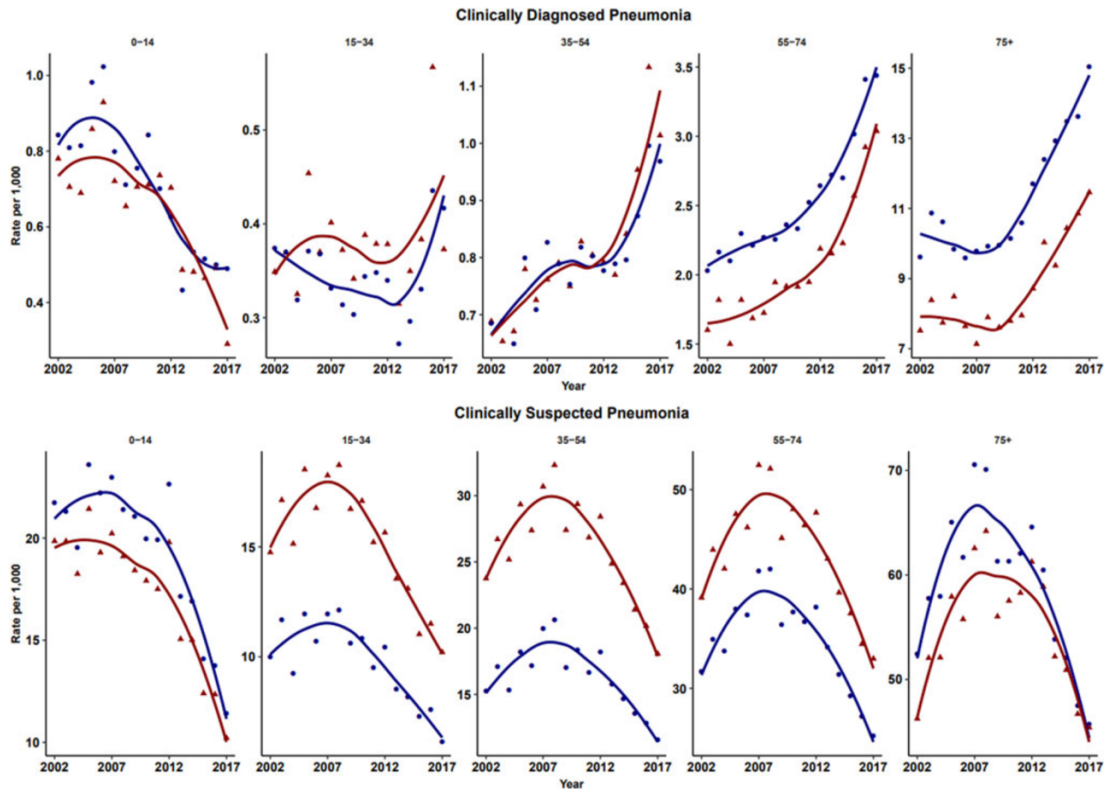


Figure 5.2: Age-specific rates for clinically-diagnosed pneumonia and clinically-suspected pneumonia for males (blue) and females (red). Rates are per 1,000 person-years

5.3.4 Sensitivity analysis

In order to evaluate whether attrition of family practices from CPRD, accounted for changes in coding, analyses were repeated using only data from 218 family practices that contributed data in every year from 2002 to 2017. In these 218 practices, the crude rate of clinically-diagnosed pneumonia increased from 1.38 per 1,000 in 2002 to 1.56 per 1,000 in 2010 and then increased to 2.24 per 1,000 in 2017 with an annual percentage change being 3.1% (1.9% to 4.2%). For clinically-suspected pneumonia, the crude rate increased from 20.8 per 1,000 in 2002 to 29.7 per 1,000 in 2008 at an APC of 4.0% (0.9% to 7.2%) before declining to 18.0 per 1,000 in 2017 at an average percentage of 5.1% (-6.6% to -3.4%).

5.4 Discussion

Main findings

There was an increasing trend in clinically-diagnosed pneumonia from 2002 onward and this accelerated after 2011. This was unlikely to be due population ageing because similar trends were observed for age-standardised and crude incidence rates. Analysis of clinically-suspected pneumonia showed that this syndrome was much more frequently recorded than clinically-diagnosed pneumonia. Its incidence increased from 2002 to 2008 but decreased rapidly thereafter. Clinicians necessarily work with diagnostic disease classifications but pulmonary infections may represent graduated phenomena with varying degrees of bronchial inflammation and alveolar consolidation. This may contribute to diagnostic uncertainties and perhaps inconsistent selection of diagnostic terms. Given that ‘chest infection’ may not be a confident diagnosis, together with the volume of antibiotic treated chest infection rates being considerably higher than that of diagnosed pneumonia, a small change in disease coding practice could lead to a shift from ‘chest infection’ recording to ‘pneumonia’ recording. Joinpoint regression analysis suggested that the decline in ‘chest infection’ recording began in 2008 slightly before the increase in ‘pneumonia’ recording from 2011. However, conditions managed as pneumonia were considered as relatively stable as none of the AAPCs have shown to be statistically significant from middle age and above. This would suggest that there was code drifting during clinical consultations when pneumonic infectious symptoms were presented among adult population with elder patient being more likely being diagnosed with pneumonia.

According to UK clinical guidelines, adult ‘chest infection’ should be managed as CAP when pneumonia is suspected (NICE, 2015). However, Petersen et al (Petersen et al., 2007a) regarded pneumonia as a potential complication of chest infection in their electronic health records based study. If antibiotics are prescribed less frequently for chest infection it is possible that pneumonia might increase (Gulliford et al., 2016). One study suggested that pneumonia might be more frequent at family practices that prescribe fewer antibiotics for respiratory infections (Gulliford et al.,

2016). Thus, although it appears likely that changes in coding of respiratory infections account for observed trends, the present data do not exclude the possibility that changing management of ‘chest infections’ is leading to an increase in pneumonia incidence. Therefore, understanding the underlying reason for adopting certain disease labels such as ‘chest infection’ rather than confident disease diagnoses during routine healthcare in the community would contribute to understanding the challenges of diagnostic uncertainty in primary care settings.

In children, records of both clinically-diagnosed pneumonia and clinically-suspected pneumonia decreased. This could be explained by the introduction of Haemophilus Influenzae Type B vaccine in 1992 and pneumococcal conjugate 7 vaccine (PCV7) in 2006 into the UK childhood immunization scheme (PHE, 2013). In adults, rates of clinically-diagnosed pneumonia increased while clinically-suspected pneumonia decreased. Trends were generally similar in males and females but women between the ages of 15 and 74 years were more likely to be recorded with antibiotic-treated chest infection than men, but this distinction was not apparent for clinically-diagnosed pneumonia. It is unclear whether this represents a disease classification preference or whether more severe cases tended to be found in men. Influenza pneumonia showed an increase in the epidemic year of 2009 (Chan, 2009) but overall low incidence of influenza pneumonia might result from the influenza vaccination schemes among both young and elder populations (PHE, 2013). Pleural infection was analysed as a surrogate of severe pneumonia because recent evidence suggests an increase in incidence rate trends among children (Mahon et al., 2016). Our results showed that the incidence trends for pleural infection remained stable during the past 16 years.

Comparison with previous studies

In contrast to previous studies on pneumonia burden using hospital admission data (Trotter et al., 2008, Quan et al., 2016b), this study using electronic health records data with GPs being data recorders. Since pneumonia patients referred to secondary care may not necessarily be representative of all community-acquired pneumonia,

using primary care consultation data contributes to understanding pneumonia patients presenting and managed in primary care settings.

In National Institute for Health and Care Excellence (NICE) guidelines for both CAP and chest infection, together with British Thoracic Society (BTS) recommendations for CAP management in adult patients, CRB-65 score (confusion, raised respiratory rate, low blood pressure and age 65 and above) is recommended to guide risk assessment and place of treatment (NICE and Excellence, 2014, NICE, 2015, BTS, 2015). This score identifies older age (≥ 65 years old) as an independent risk score since being 65 years and above will automatically classify patients into an intermediate-risk group. This implies that more conservative management strategies have been applied to older pneumonia patients. This partially explained previous study findings that pneumonia patients in the community leading to hospitalization were increasing in recent years among elder populations and there is no evidence that less severe patients were admitted to the secondary care.

Strengths and limitations

This population-based study analysed healthcare records to outline the range of conditions that were managed as pneumonia in the community. The 16-year timeframe provided sufficient data to estimate disease trends over a substantial study period. The study included all eligible practices and patients contributing health care information to CPRD more than one year. There were 320 practices participating the database by the end of 2017. Previous studies have shown the completeness and data of high quality in CPRD. The large sample size of research cohort was sufficient for depicting the trends for clinically-diagnosed pneumonia, clinically-suspected pneumonia, even low incidence conditions like influenza pneumonia and pleural infections.

The chronological trends in this study were mapped through joinpoint regression analysis which allows the estimation of changing point generated from the model

rather than set arbitrarily. By identifying the changing time points, possible effects of influential health policies or public health measures could be investigated especially when the starting date is easily established. Also, the quantified average change during study period would provide complementary information to interrupted time series analysis for intervention evaluation (Kontopantelis et al., 2015). In this study, clinical diagnosed pneumonia incidence was found to increase around 2010 to 2011 suggesting conservative antibiotic prescribing strategies were applied in primary care. This may result from various antibiotic stewardship initiatives such as ‘Start Smart- Then Focus’ in the UK healthcare system (Ashiru-Oredope et al., 2012a). Similar application was reported in the context of coronavirus 2019 (Covid-19) to investigate the effectiveness of national lockdown (Chaurasia, 2020a).

Free text information documented in CPRD was not available for this study (Wolf, 2018), and such information would enable us to examine diagnostic information documented in free text records rather than coded. But we consider that clinicians would record relevant conditions like pneumonia or chest infection especially when clinical discretion leads to antibiotic treatment. Also, GPs might have shifted to codes that were not included in our case definition such exacerbations of COPD that could be previously diagnosed as chest infection. Further, health care information in CPRD made disease severity assessment unfeasible, therefore, we could not confidently determine whether case severity influenced coding practices or place of treatment. Plus, we did not capture data from out-of-hour services, walk-in centre consultations and emergency care.

This study was derived from a universal health care coverage system where most common conditions are managed in primary care settings, implications generated from this study would mainly apply to similar systems but not the ones where health care insurance plays an essential role or referral thresholds from primary care to secondary care vary extensively compared to that in the UK. Practice variability was implied in the results but not explored as it was not a main focus of this study. Caution is needed when generalising research findings to the whole primary care system in the UK.

5.5 Conclusion and implication for future research

Clinically-diagnosed pneumonia is increasing over time in the UK. This trend could not be fully explained by aging population, changes in coding practice or alternative diagnosis. Divergency was found in age-specific trends with decreasing trends in children but increasing in older adults. Respiratory conditions managed as pneumonia in family practice were decreasing slightly over time, which was more likely due to more conservative antibiotic prescribing strategies. Research to reduce diagnostic uncertainty would contribute to improving antibiotic stewardship in the community.

Chapter Six : Prognostic factors for development of a model in community-acquired pneumonia in adults: a systematic review

A systematic review of literature was conducted to identify prognostic factors for incident CAP in adults. This chapter presents a brief introduction followed by the study methods, results and conclusions.

6.1 Introduction

As described in previous chapters, current clinical guidelines for pneumonia management in primary care rely on evidence mostly derived from cohort studies which investigated severe CAP, with hospital admission, emergency hospital visits or mortality as outcomes. Risk factors for severe pneumonia, including pneumonia requiring urgent or intensive medical care as well as mortality, may not continue to apply to all CAP cases. In community settings, many patients are at low risk of developing complications or undesirable treatment outcomes. However, prompt identification of CAP in patients who present with early symptoms in primary care may contribute to timely treatment and appropriate clinical investigation if this differential diagnosis is considered, and this may eventually lead to better healthcare outcomes. Therefore, prognostic factors identified through prediction studies will contribute to identifying modifiable risk or protective factors and stratified patient management (Hayden et al., 2006). Prediction modelling should take both clinical and statistical significances into consideration. Inclusively reviewing relevant potential prognostic factors following a systematic approach will assist in developing the prediction modelling process, including predictor identification, predictor sorting, study design and statistical modelling approach selection, as well as informing explanation of results (Altman, 2001).

The question for this review was formulated using the 'PICOS' framework (('P': patient problem or population; 'I': interventions; 'C': comparators; 'O': outcome(s) and 'S': study design), which is commonly adopted for constructing research

questions as an evidence-based framework (CRD, 2009). Furthermore, the research question was refined after referencing the 'FINER' criteria, which capture the characteristics of a good study question which are: feasible, interesting, novel, ethical and relevant (Hulley et al., 2001). The key components of review questions population, intervention, comparator and outcome are explained below with study design outlined in the methods section.

Population

For the prediction modelling in this thesis, the study was conducted among the general adult population with RTI consultations preceding to pneumonia diagnosis within 30 days. However, the study population for this systematic review was expanded to include the general adult population in the community. This is because risk factors for incident CAP reported by existing literature still have relevance for this study population. That is, the prognostic study population could be regarded as a subgroup of general adult population. Also, if the study cohort was restricted to patients consulting within RTI only, then a limited number of studies would be expected to be eligible making results less informative. Therefore, this systematic review was conducted among general adult population or adult population with common comorbidities including COPD or diabetes.

Interventions and comparators

Both risk and protective factors that would contribute to increase or decrease the risk of incident CAP were included for review. Clinical investigative tests that could facilitate prompt diagnosis of CAP were also included. However, effects for well-established therapeutic interventions such as antibiotics may not always be explored through experimental studies since it would be unethical to hold back the treatment when there are evident clinical indications.

Outcomes

The endpoint of the prediction modelling of this thesis is defined to be adult incident CAP after RTI consultations during preceding 30 days. CAP comprises two essential concepts in this scenario: the environment in which patients acquire pneumonia is a community setting in which empirical antibiotic treatment is generally recommended; further CAP denotes that it is a broad and common clinical condition with an array of common causal pathogens that would normally respond to the recommended initial empirical treatment plan, rather than any specific atypical pneumonia (Olson and Davis, 2020, Mackenzie, 2016). Meanwhile, any subtype of non-infectious origin pneumonia like bronchiolitis obliterans organizing pneumonia (BOOP) which antibiotic treatment does not constitute as an essential effective management was not covered by CAP (Epler, 2001, Al-Ghanem et al., 2008). In this systematic review, the case definition of incident CAP was accepted as originally defined in each individual study and critically appraisal was conducted based on study results for sensible inferences. Admittedly, the inclusion of a more homogenous case definition may be beneficial in terms of minimizing misclassification and bias, but application of stringent case definition may not be able to answer the review question in a generalisable or transferrable manner especially, for research studies conducted in primary care where confirmatory diagnostic tests are not always available.

Conducting a systematic review made it possible to account for the heterogeneity between individual primary prognostic studies, in terms of study participant characteristics, study design, case definition, methodological quality, prognostic factor selections and confounding variable measurement and biases (Hemingway et al., 2009, Hayden et al., 2009). No attempt was made to synthesise the measures of association of prognostic factors with CAP through formal meta-analysis. Both primary prognostic studies and systematic reviews of prognostic factors or intervention studies were included to identify potential prognostic variables for incident CAP. Further, individual primary prognostic studies were evaluated for methodological quality. Evidence for prognostic factors was combined in a narrative synthesis where the overall risk of bias was low (Hayden et al., 2006).

Based on previous study of pneumonia incidence trends in the UK general population, clinically diagnosed pneumonia incidence rates have shown to have an increasing chronological trend with acceleration after 2010 (Sun et al., 2019). Therefore, this systematic review included studies published since 2010 so that contemporary information could be summarized to inform model development, but evidence generated from earlier studies could be included through previous systematic review evidence that included the earlier period.

6.2 Objectives

The main objectives were to:

1. Identify possible candidate predictors from previous studies of sufficient quality to inform variable selection during prediction modelling;
2. Scope the statistical modelling methodology deployed by previous prognostic studies of incident CAP;
3. Critically appraise existing evidence especially when conflicting results were presented, to explain where the possible reasons derived i.e. study design, variation in methodology or simply by random chance.

6.3 Methods

The review was guided by the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) (Moher et al., 2010) and PRISMA Extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018).

6.3.1 Study selection

Studies were selected according to the following criteria:

Inclusion criteria

Observational studies including cohort studies (both prospective and retrospective), cross sectional studies, case-control studies, interventional studies (both randomized controlled trials and quasi-experimental studies) and systematic reviews which

reported the association between prognostic factors and adult incident CAP were included. Case definition of CAP was accepted as originally defined in each study. Studies of pneumonia that did not explicitly identify CAP, but the patient population indicated a community setting or its equivalent i.e. outpatient clinic, were also included for review. Study results with subgroup analyses that met the inclusion criteria were also included in this review. Articles with English full text published between 1st January 2010 and 20th January 2020, at the latest, were eligible for inclusion.

Exclusion criteria

Studies conducted among children younger than 16 years old were excluded. Laboratory based biological studies, case reports, surveys, editorials, qualitative studies, reviews that did not follow systematic approaches and research on CAP that were caused by specific atypical, or drug resistant causal pathogens were not reviewed for inclusion. Prognostic studies with outcomes of interest being recurrent CAP, CAP severity, CAP hospitalization, CAP mortality or CAP treatment effect evaluation were excluded. For study cohorts being CAP patients treated in hospital setting including emergency departments and inpatient wards or nursing home dwellers were excluded from this systematic review. Also, studies conducted in patient cohorts with specific non- prevalent medical conditions like HIV infections were not included for final review.

Selection criteria

We exported citations from databases into EndNote X9 (Analytics, 2018) for de-duplication and screening. The author screened the titles and abstracts against the inclusion and exclusion criteria. Full text articles of potentially eligible studies were independently reviewed by the author (XS) and the lead supervisor (MG). Conflicts were resolved through consensus.

6.3.2 Search strategy

The protocol for this systematic review was registered on the International Register of Systematic Review (PROSPERO) (registration number: CRD42020168684). Databases were searched for published (MEDLINE, Embase, Cochrane Database of Systematic Reviews and Cochrane Central Register of Controlled Trials (CENTRAL)) and unpublished (OpenGrey) studies. Citations and references of included studies were screened to identify additional studies that might not have been captured during database searches. Searching strategy was developed by referencing previous systematic reviews in respiratory infection (McDonagh et al., 2016, Chalmers et al., 2010) and guidance on prognostic factor studies (Riley et al., 2019a, Altman, 2001). Search strategies for Medline and Embase are displayed in Appendix A. For Cochrane Database of Systematic Reviews, CENTRAL and OpenGrey, search terms only included 'pneumonia' and 'community', so that all potentially relevant studies were included.

6.3.3 Data extractions

Data extraction was performed by XS into pre-designed tables. All articles included for full text analysis were also reviewed by MG. Key components of data extraction included the title, author, publication year, country/ region, study population, number of participants, sex composition, age range, study design, statistical modelling methodology, identified prognostic factors, prediction end point. Extracted data were reviewed by MG for accuracy. The final included studies as agreed by both reviewers, were assessed for further synthesis.

6.3.4 Quality assessment

Included studies were evaluated independently by two reviewers (XS and MG). Individual prognostic studies other than systematic reviews were assessed using the Quality in Prognosis Studies (QUIPS) tool that assesses risk of bias for prognostic factor studies in six domains: participation, attrition, prognostic factor measurement, outcome measurement, confounding, and statistical analysis and reporting (Hayden

et al., 2013). Disagreements in bias assessment were addressed through discussion and consensus between XS and MG. An overall level of bias risk was assigned to each prognostic study. If anyone of these six domains was considered to be at high risk of bias, an overall high risk of bias was given. An overall moderate risk of bias was assigned to studies when none of the six domains was at high risk of bias meanwhile three or more domains were evaluated to be at moderate risk. Studies were assigned an overall low risk of bias if three or more domains were considered to be at low risk and no high risk domain was identified (Hayden et al., 2006, Burton et al., 2016).

6.3.5 Data analysis

Study characteristics were reported as counts and proportions. Heterogeneity was examined through reviewing extracted information on study population, study design, prognostic factor measurement, statistical methodologies and outcome definitions. There was variation in eligibility criteria, case definition, study design, study population, prognostic variable measurements, statistical approaches to adjustment for covariates and outcome definition across included studies. Therefore, association between prognostic factors and incident CAP was explored using a narrative synthesis approach (Sheehan et al., 2018, Altman, 2001). Evidence of prognostic factors from individual prognostic researches with an overall low risk bias was synthesised (Hayden et al., 2006).

6.4 Results

6.4.1 Study selection

Figure 6.1 presented study selection process of this systematic review. 10,293 references were identified after de-duplication, out of which 10,137 were excluded by titles and abstracts. 156 studies were reviewed by full text for intended inclusion with 110 articles failed to meet inclusion criteria as shown in Table A 9. Therefore, 46 studies were included for narrative synthesis and risk of bias assessment was performed subsequently among 30 of these included studies (Table 6.5).

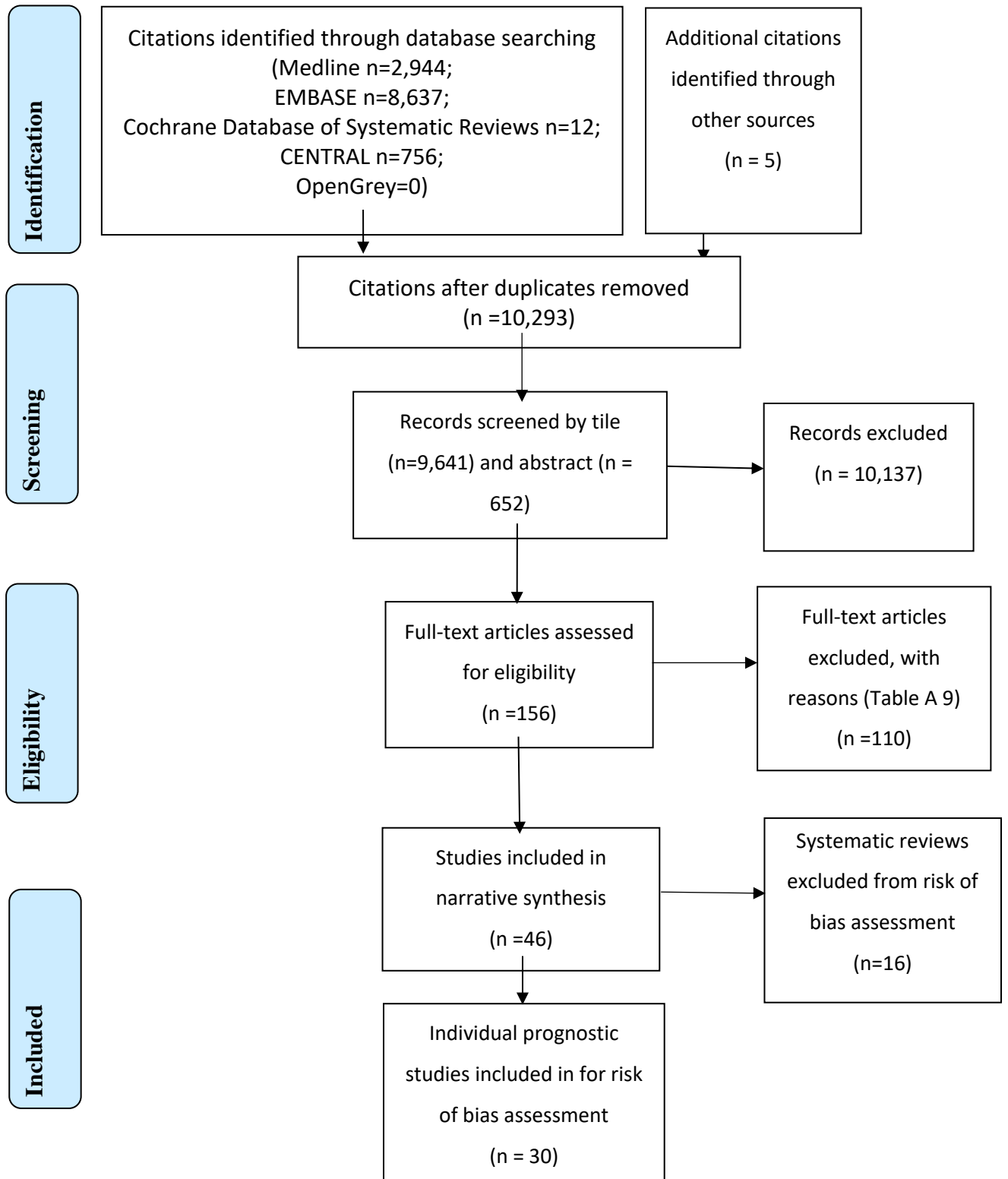


Figure 6.1: PRISMA flow diagram outlining systematic review process

6.4.2 Study characteristics

This systematic review has included 30 prognostic studies with 3,432,436 patients (the mean/ median age for different studies ranged from 43 to 83 years old) and 16 systematic review studies with sample sizes ranging from 1,214 to 6,546,396. Most studies were conducted in European countries (Table 6.1 and Table 6.2). All individual prognostic studies have adopted multivariable approaches with regression based statistical models to adjust confounding variables and quantify effects sizes. Apart from 3 systematic reviews, meta-analysis has been deployed to synthesis quantitative evidence in the remaining 13 systematic reviews.

6.4.3 Prognostic factors

In total, there were 33 prognostic factors for incident CAP identified by 46 included studies (Table 6.3 and Table 6.4) which could broadly categorized into: patient general impression (older age, respiratory tract infection symptoms), biomarkers (CRP, procalcitonin (PCT), 25(OH)D levels), use of medication (PPI, statin, steroids, anticholinergics, β -blockers, calcium channel blockers (CCBs), Angiotensin-converting enzyme (ACE) inhibitors, thiazide, antibiotics, anticholinergic, antipsychotic, diabetic drugs, immunosuppressant drugs), life style and environmental exposures (physical activity, functional status, body weight, smoking/ passive smoking, dental hygiene, season (winter)), preventive procedures (flu vaccination, pneumococcal vaccinations (PPV23, PCV13)), comorbidity and underlying health conditions (cardiopulmonary conditions, renal function impairment, lung cancer, HIV and previous CAP history). In order to provide a simplified summary of the complex data, variables with their risk or protective effects on incident CAP concluded being statistically significant after adjusting for other covariables or cofounding variables were labelled as 'Yes', otherwise as 'No' as shown in Table 6.3 and Table 6.4.

There were conflicting results identified concerning several predictive effects of certain variables including CRP, weight gain/ obesity, flu and pneumococcal vaccinations. Effect estimates of prognostic factors that were identified from

individual prognostic studies rated as being at low risk of bias were organized further in section 6.4.4.

6.4.4 Quality assessment for individual prognostic studies

The results of the quality assessment results of individual prognostic studies in this systematic review were presented in Table 6.5. In terms of the risk of bias domains, 95% agreement was reached between two independent reviewers and 100% consensus was attained following discussions. The final results of quality assessment of these 30 individual prognostic studies were presented in Table 6.5. There were 40% (12 out of 30) of individual studies that were rated as low risk of bias, 13.3% (4) were rated as moderate risk of bias with the remaining 14 studies (46.7%) considered as high risk of bias.

Among the high risk of bias studies, there were 4 studies ranked as high bias mainly due to statistical analysis approaches or inadequate results reporting. Another 4 studies were due to lack of control of study confounding variables. Study participation bias was introduced in five studies. One study was rated high bias in prognostic factor and outcome measurement. For study attrition, no study was ranked with high bias in this domain.

Predictive factors with direction and effective size quantified by individual low risk of bias were presented in Table 6.6 including elder age, smoking, poor dental hygiene, the use of PPI, ICS, Oral antidiabetic agents (OADs) in combination with thiazolidinedione, anticholinergic medication, statin and antipsychotic medication, COPD, asthma, bronchiectasis, kidney function impairment, lung cancer and HIV. Protective effect of PCV 13 against CAP remained to be less conclusive.

Table 6.1: Characteristics of included studies

Author	Year	Country/ Region	Study population	No. of participants	Sex composition	Age range
Gessner et al.	2019	the Netherlands	Elder adult (65 and above)	84,496 in total: PCV13 group: 42,237; placebo group: 42,255	PCV13 group: 44.5% Female, 55.5% Male; placebo group: 43.7% Female, 56.3% Male	PCV13 group: 72.8±5.7; placebo group: 72.8±5.6
Kolditz et al.	2019	Germany	Patients 60-99 years old who received a PCV-13 vaccination between 1 January 2012 and 31 December 2016 and no other pneumococcal vaccination within the above-specified time frame.	Cases: 11,395; controls: 34,185	Case: 58.6% Female; 41.4% Male;	Median age: 75 (IQR: 67-82)
Rivero-Calle et al.	2019	Spain	Adults aged 18 and older	2,332,622 participants	48.3% Female, 51.7% Male	Mean age 60.5 (SD: 20.3)
Zirk-Sadowski et al.	2018	England	Adults (60 and above)	Cases: 75,050; Controls: 75,050	All participants 58% Female;	Mean age for all participants: 71(±7.3)
Janson et al.	2018	Sweden	Adults (40 and above)	Cases: 6,623; Controls: 48,566	Cases: 55.7% Female, 44.3%; Male:	Cases: mean age 65.9 (± 10.1);
Gorricho et al.	2017	Spain	Adult type-2 diabetes (T2DM) patients	Cases: 1,803; Controls: 17,986	Case: 36.9% Female, 63.1% Male; Controls: 36.9% Female, 63.1% Male	Cases: mean age: 71.7 (SD: 12.4); Controls: mean age: 71.6 (SD: 12.3)
Paul et al.	2015	USA (Washington State)	Adults aged 65 to 94	Cases: 1,039; Controls: 2,022	49% Female and 51% Male for both cases and controls	Median age 77 (IQR: 71-82) for both cases and controls

McDonald et al.	2015	UK	Adults 65 years or older with diabetes mellitus and no history of renal replacement therapy (Patients without creatinine results were excluded.)	191,709 patients	53% Female, 47% Male.	Median age at study entry was 71 (IQR: 66-78) years.
McKeever et al.	2013	UK	Adults aged 18 to 80 years old	6,857 cases and 36,312 controls	Cases: 60.9% Female, 39.1% Male; Controls: 62.3% Female, 37.7% Male	Cases: mean age: 55.5 (SD: 17.8); Controls: mean age: 53.7 (SD: 17.9)
Lin et al.	2013	Taiwan	Newly diagnosed COPD patients based on pulmonary function test without asthma	2,630 patients	CAP patients: 21.1% Female, 78.9% Male; 15.9% Female, 84.1% Male.	CAP patients: median age: 70.1 years old (IQR: 58.9-78.1); Non-CAP patients: median age: 77.3 (IQR: 68.6-82.6)
Vinogradova et al.	2011	UK	Patients age 45 and above	17,755 cases and 80,484 controls	Cases: 51.3% Female, 48.7% Male; Controls: 51.2% Female, 48.8% Male	Median age: 74 (IQR: 62-82);
Trifiro et al.	2010	the Netherlands	Adults aged 65 and older	258 cases and 1,689 controls	Cases: 55% Female, 45% Male; Controls: 72.9% Female, 27.1% Male	Cases: mean age: 83.6 (SD: 7.4); Controls: mean age: 83.2 (SD: 5.9)
Moore et al.	2019	The UK	Adult uncomplicated LRTI patients	28,883 adult patients	NA	All patients were 65 and above
Williams et al.	2017	UK	COPD patients aged 40 and above	14,513	46.6% Female, 53.6% Male	Mean age 70.3 (\pm 10.8)
Othman et al.	2016	UK	Adults aged 18 and older	320,000	55% Female; 45% Male	Mean age: 56 (SD: 16)

Steurer et al.	2011	Switzerland	Patients 18 years and above who presented with cough or worsened cough together with increased body temperature, patients with known chronic lung disease (except chronic bronchitis) were excluded	127 cases, 494 non-cases	50% Female, 50% Male for both cases and non-cases	Mean age: 46.8 (SD: 16.3)
Yeh et al.	2019	Taiwan	Adult (20 and above)	CVD group: 28,363; control group: 28,363	48.5% Female; 51.5% Male	CVD group (mean age 49, SD: 13.5); Non-CVD group (mean age: 49.2, SD: 13.1)
Groeneveld et al.	2019	the Netherlands	Adult acute RTI patients	249	51.0% Female; 49% Male	Median age 56 (IQR: 43-67)
Rivero-Calle et al.	2016	Spain	Adults aged 18 and older	2,332,622 participants	48.3% Female, 51.7% Male	Mean age 60.5 (SD: 20.3)
Almirall et al.	2014	Spain	Participants aged 14 and above	1,336 cases and 1,326 controls	60% Female, 40% Male	Mean age: 59.6 years old (SD: 20.0)
van Vugt et al.	2013	12 European countries	Adults presenting with acute cough	2,820 participants	60% Female, 40% Male	Mean age: 50
Quraishi et al.	2013	USA	Cross-sectional sample of the non-institutionalized civilian population from the Third National Health and Nutrition Examination Survey (NHANES III) who were 17 and older	16,975 participants	62.7% Female; 37.3% Male	Median age: 43 (IQR: 29-64)
Vila-Corcoles et al.	2012	Spain	Patients diagnosed with chronic pulmonary conditions (chronic bronchitis, emphysema and/or asthma) aged 50 and older	96 cases and 192 controls	16.7% Female, 83.3% Male	Cases: mean age: 73.2 (SD: 10.4); Controls: mean age: 72.9 (SD: 9.7)

Mullerova et al.	2012	UK	COPD patients aged 45 and above	40,414 COPD patients aged 45 and above	48.3% Female, 51.7% Male	45-64: 37.1%; 65-79: 49.9%;
Almirall et al.	2010	Spain	Participants aged 14 and above	1,336 cases and 1,326 controls	60% Female, 40% Male	Mean age: 59.6 years old (SD: 20.0)
Mukamal et al.	2010	US, Puerto Rico, and the US Virgin Islands.	81,000 middle age individuals used at least one antihypertensive medication	7,429 cases and 73,571 controls	56% Female, 44% Male	Cases: mean age: 58.4 (± 12.8); Controls: mean age: 58.2 (± 12.5)
Neuman et al.	2010	US	Women aged 25 to 42 years old who did not report pneumonia at base line without conditions including cancer, cardiovascular disease (myocardial infarction, stroke, or arterial surgery), or asthma during study period.	83,165 women with 965,168 person-years during 12 years follow up (1,265 incident CAP)	100% Female	Age ranging from 25 to 42
Jackson et al.	2009	USA (Washington State)	Immunocompetent senior adults (aged 65 to 94)	Cases: 1,173; Controls: 2,346	49% Female and 51% Male for both cases and controls	38% were younger than 75, 45% were aged 75 to 84, and 17% were aged 85 and older.
Baik et al.	2000	USA	Health professionals	104,491 participants	74.7% Female, 25.3% Male	Female: 27-44 years old; Male: 44-79 years old
Evertsen et al.	2010	US	Adult patients aged 18 to 80 years old	Pneumonia: 4,907; bronchitis: 32,760; URTI: 20,037	Pneumonia: 52.0% Female, 48.0% Male; bronchitis: 62.4% Female, 37.6% Male; URTI: 61.6% Female. 38.4% Male	Pneumonia: 45.9 \pm 13.4; bronchitis: 40.4 \pm 12.2; URTI: 40.4 \pm 12.6
Marchello et al.	2019	NA	Adult	NA	Both	NA

Baskaran et al.	2019	NA	Adult of all age	NA	NA	NA
Zhou et al.	2019	NA	Ranging from 3 days to 94 years old	20,966 participants	NA	Ranging from 3 days to 94 years old
Htun et al.	2019	Iran (n = 1), USA (n = 3), Denmark (n = 2), Netherlands (n = 2), Norway (n = 2), Sweden (n = 1), Switzerland (n = 1) and Europe (n = 1)	adults (18 and above)	11,144 participants	NA	Adults (18 and above)
Baskaran et al.	2018	NA	Adult	460,592	NA	NA
Walters et al.	2017	NA	COPD patients	2,171 participants from 12 RCT	All participants: 33% Female; 67% Male	Average 66 years old
Htar et al.	2017	NA	Adult general population, immunocompromised and subjects with underlying risk factors	NA		
Almirall et al.	2017	NA	14 and above	More than 169,018	NA	NA
Schiffner-Rohe et al.	2016	Europe and Asia	Adults aged 60 and older	30,171 participants	NA	Adults aged 60 and older

Lambert et al.	2015	Spain, Taiwan, USA, UK, Canada, Denmark, the Netherlands, Italy, Australia, Europe, Japan	Adult participants	6,546,396 participants	NA	NA
Abramowitz et al.	2015	South Korea, Canada, USA	Mixed between children and adults	NA	NA	NA
Phung et al.	2013	Australia, Demark, Korean, Spain, Japan, German, USA, UK, Canada, China, Singapore	Mixed between children and adults (12-87 years old)	2,561,839 participants	NA	NA
Khan et al.	2013	UK, US and Canada	Participants age range from 15 and above	NA	NA	Participants' age range from 15 and above
Loeb	2010	NA	4,482 older people residing in care homes; 814 older people residing in care homes; 18,090	NA	NA	NA

			older people living in the community			
Johnstone et al.	2010	Canada and Europe	Adult patients aged 18 and over	Approximately 1 million patients were involved	Female: 47%~59%	NA
Engel et al.	2012	Three studies from Denmark ¹ , the Netherlands ² and UK ³	Adults with LRTIs (1& 2: 18 years old and above; 3: 16 years old and above)	1,214	NA	Adults ranging from 16 years old and above

Table 6.2: Characteristics of included studies (continued)

Author	Year	Study design	Modelling/ Statistical methodology	Identified risk factors	Prediction end point
Gessner et al.	2019	Parallel-group, double-blind, placebo controlled clinical trial (13-valent pneumococcal conjugate vaccine (PCV13) vs placebo) with the mean and median years of follow-up per subject varied from 3.9 to 4.0 for both trial arms	Modified intention-to-treat (mITT) analysis	PCV13 was effective in preventing vaccine-type pneumococcal, bacteremic, and nonbacteremic CAP and vaccine type invasive pneumococcal disease but not in preventing community-acquired	Clinical incident CAP, pneumococcal CAP (SpCAP) and vaccine-type pneumococcal (VT-Sp) CAP
Kolditz et al.	2019	Retrospective cohort study	Multivariable logistic regression analysis	PCV-13 vaccination in adults ≥ 60 years was associated with a significant risk reduction of all-cause pneumonia.	All cause pneumonia
Rivero-Calle et al.	2019	Retrospective cohort study	Binary logistic REGRESSION	CAP risk increases with age and doubles in males older than 75 years, comorbidities associated with CAP were metabolic disease, cardiovascular disease, and diabetes.	Incident CAP
Zirk-Sadowski et al.	2018	Retrospective nested case-control study (matched on age and sex)	Cox proportional hazard models	Proton-Pump Inhibitors (PPI)	Incident CAP

			adjusted by propensity scores		
Janson et al.	2018	Retrospective nested case-control study (matched on age, gender, and the starting year of the index date of COPD)	Cox regression model	COPD, asthma and ICS use	Incident CAP
Gorricho et al.	2017	Retrospective nested case and control study (matched on age, sex and calendar year 1:10)	Conditional logistic regression	Thiazolidinedione use in combination was associated with an increase in the risk of CAP when compared to metformin + sulfonylureas. The use of DPP-4 inhibitors was not associated with an increased risk of CAP.	Incident CAP
Paul et al.	2015	Nested case and control study (matched on age, sex and year with 1:2 ratio)	Conditional logistic regression	Anticholinergic medication use is associated with CAP risk among elder adults.	CAP
McDonald et al.	2015	Retrospective cohort study	Poisson regression with lexis expansions for age and a random-effects model.	Both eGFR and proteinuria were independent risk markers for incidence of pneumonia among elder diabetic patients.	Incident CAP as subset of LRTIs
McKeever et al.	2013	Nested case and control study (matched on age, sex and	Conditional logistic regression	People with asthma receiving inhaled corticosteroids are at an increased risk of	CAP

		index date (within 3 years) with 1:6 ratio)		pneumonia with those receiving higher doses being at greater risk.	
Lin et al.	2013	Retrospective cohort study	Cox's proportional hazards regression models	The likelihood of CAP increased	Incident CAP
Vinogradova et al.	2011	Nested case and control study (matched on age (within 1 year), sex, practice and calendar year)	Conditional logistic regression	Current exposure to statins was associated with a reduced risk of pneumonia.	Incident CAP
Trifiro et al.	2010	Nested case and control study (matched on year of birth, sex and index date)	Conditional logistic regression	The use of either atypical or typical antipsychotic	Incident CAP
Moore et al.	2019	Retrospective cohort study	Generalized linear model of binomial family	symptom severity (absence of coryza, fever, chest pain,	Serious adverse outcomes (late-onset pneumonia,
Williams et al.	2017	retrospective cohort study	Multivariate logistic regression	Older age, increasing grade of airflow limitation, lower body mass index, inhaled corticosteroid use, prior frequent exacerbations, comorbidities, (including ischemic heart disease and diabetes) and winter	Incident CAP and AECOPD

Othman et al.	2016	Nested case and control (cases and controls were matched on age (within 5 years), sex, and year of PPI prescription with 1:1 ratio)	Cox model and conditional Poisson regression models with fixed effects with adjustment for age (5-year bands)	The association between the use of PPIs and risk of CAP is likely to be due entirely to confounding factors.	CAP (pneumonia and pneumonia plus chest infection, LRTIs)
Steurer et al.	2011	Prospective cohort study	Multiple logistic regression model and CART for modelling. Multiple imputation for missing value	In patients with C-reactive protein values below 10 µg/ml or patients presenting with C-reactive protein between 11 and 50 µg/ml, but without dyspnoea and daily fever, pneumonia can be ruled out.	CAP
Yeh et al.	2019	Retrospective cohort study	Cox proportional hazard regression models; propensity score was used for matching	Pneumonia risk was associated with CVDs, especially heart failure, regardless of age, gender, comorbidities, and antibiotic use, particularly in elderly male patients.	Incident CAP

Groeneveld et al.	2019	Prospective observational cohort study	Multivariate binary logistic regression	Patient feeling ill and absence of runny nose are predictive for CAP onset whereas CRP predicts pneumonia better than PCT and	CAP onset
Rivero-Calle et al.	2016	Retrospective cohort study	Binary logistic regression	CAP risk increases with age and doubles in males older than 75 years, comorbidities associated with CAP were metabolic disease, cardiovascular disease, and diabetes.	Incident CAP
Almirall et al.	2014	Nested case and control study (matched on age (within 5 years), sex and primary healthcare area with 1:1 ratio)	Logistic regression	Inhaled steroids may favour CAP in COPD patients, whereas anticholinergics may favour CAP in asthma patients. In chronic bronchitis (CB) patients, no association with CAP was observed for any inhaler.	Incident CAP verified via chest radiography
van Vugt et al.	2013	Cross sectional observational study	Multivariable logistic regression for model development, boot strapping for internal validation	Absence of runny nose and presence of breathlessness, crackles and diminished breath sounds on auscultation, tachycardia, and fever. CRP>30mg/L were associated with increased risk of pneumonia.	CAP determined by radiographs
Quraishi et al.	2013	Cross-sectional study	Multivariable logistic regression, locally weighted scatter plot smoothing	25(OH)D levels were inversely associated with history of CAP.	CAP

			(LOWESS) to examine the association between 25(OH)D level and the cumulative frequency of CAP.		
Vila-Corcoles et al.	2012	Nested case and control study (matched on primary care centre, age, sex, and main comorbidity)	Conditional logistic regression	The effectiveness of the PPV-23 in preventing pneumonia among patients with chronic pulmonary disease is uncertain.	CAP
Mullerova et al.	2012	Population-based retrospective cohort study	Multivariate conditional logistic regression	Age over 65 years was significantly associated with increased risk of CAP. Other independent risk factors associated with CAP were co-morbidities including congestive heart failure and dementia. Prior severe COPD exacerbations requiring hospitalization and severe COPD requiring home oxygen or nebulised therapy were also significantly associated with risk of CAP.	Incident CAP
Almirall et al.	2010	Nested case and control study (matched on age (within 5 years), sex and primary healthcare area with 1:1 ratio)	Logistic regression	Inhaled steroids may favour CAP in COPD patients, whereas anticholinergics may favour CAP in asthma patients. In chronic bronchitis (CB) patients,	Incident CAP verified via chest radiography

				no association with CAP was observed for any inhaler.	
Mukamal et al.	2010	Nested case and control study (matched on age (within 1 year), sex, US Census region of residence, insurance plan, subscriber status (insured individual, spouse, or dependent) with 1:10 ratio)	Conditional logistic regression	Risk of pneumonia was higher among users of β -blockers, calcium channel blockers and lipophilic ACE inhibitors in the preceding 3 months; risks were also higher for use in the preceding 12 months. Lower risk was observed among thiazide users in the preceding 3 months.	CAP
Neuman et al.	2010	Prospective cohort study	Cox proportional hazards multivariate models	Higher physical activity does not substantially reduce pneumonia risk in well-nourished women.	Incident CAP
Jackson et al.	2009	Nested case and control study (matched on age (within one year) and sex with 1:2 ratio)	Conditional logistic regression	Immunocompetent senior adults with cardiopulmonary disease, poor functional status, low weight, or recent weight loss have a greater risk of developing CAP.	CAP
Baik et al.	2000	Prospective cohort study	Multiple logistic regression	Smoking and excessive weight gain are risk factors for CAP among men and women, and physical activity was inversely associated with risk of CAP only among women.	CAP

Evertsen et al.	2010	Retrospective cohort study	Multivariate logistic regression	The presence of abnormal breath sounds and a temperature > 100°F were the predictors of a pneumonia diagnosis.	Incident CAP
Marchello et al.	2019	Systematic review	Bivariate meta-analysis	normal vital signs (temperature, respiratory rate, and heart rate) and normal pulmonary examinations	Non- CAP cases
Baskaran et al.	2019	Systematic review	Random-effects meta-analysis	Tobacco smoke exposure is significantly associated with the development of CAP in current smokers and ex-smokers. Adults aged > 65 years who are passive smokers are also at higher risk of CAP.	CAP
Zhou et al.	2019	Systematic review	Random-effects and/ or fix-effect meta-analysis	D deficiency increased the risk of CAP	CAP
Htun et al.	2019	Systematic review	Random effect meta-analysis	clinical features including respiratory rate $\geq 20 \text{ min}^{-1}$ (3.47; 1.46–7.23), temperature $\geq 38^\circ \text{C}$ (3.21; 2.36–4.23), pulse rate $> 100 \text{ min}^{-1}$ (2.79; 1.71–4.33), and crackles (2.42; 1.19–4.69); PCT $> 0.25 \text{ ng/ml}$ and CRP $> 20 \text{ mg/l}$ were predictive for CAP	CAP
Baskaran et al.	2018	Systematic review	Random effect meta-analysis	Current and ex-smoker were more likely to develop CAP compared to never smoker; passive smoking is a risk factor for people age 65 and above; higher tobacco consumption had higher risk of CAP among current smokers	CAP

Walters et al.	2017	Systematic review	Random-effects and/ or fix-effect meta-analysis	pneumococcal conjugated vaccine (PCV) had protective effects on CAP, but such protective effect on VT CAP was not supported by sufficient evidence.	CAP
Htar et al.	2017	Systematic review	Random-effects model	PPV23 against CAP was not consistent in the general population, the immunocompromised and subjects with underlying risk factors	Any CAP, Pneumococcal CAP, non-bacteremic-pCAP,
Almirall et al.	2017	Systematic review	Random-effects model	Age, smoking, environmental exposures, malnutrition, previous CAP, chronic bronchitis/chronic obstructive pulmonary disease, asthma, functional impairment, poor dental health, immunosuppressive therapy, oral steroids, and treatment with gastric acid-suppressive drugs were definitive risk factors for CAP.	CAP and hospitalized CAP
Schiffner-Rohe et al.	2016	Systematic review	Random-effects model	No proof that PPV23 can prevent pCAP or CAP in a general, community-dwelling elderly population.	All cause CAP (pCAP as subgroup endpoint)
Lambert et al.	2015	Systematic review	Random-effects model	Outpatient PPI use is associated with a 1.5-fold increased risk of CAP	CAP

Abramowitz et al.	2015	Systematic review	Quantitative information was not synthesized	CAP is noted to associated with shorter duration of therapy	CAP
Phung et al.	2013	Systematic review	Random-effects meta-analysis for pooled effect sizes; two-order fractional polynomial and random-effects meta-regression analysis for BMI-pneumonia dose-response effect	A J-shaped relationship between BMI and risk of CAP (underweight, RR 1.8, 95% confidence interval [CI], 1.4–2.2, P < 0.01; overweight, 0.89, 95%CI, 0.8–1.03, P, 0.1; obesity, 1.03, 95% CI, 0.8–1.3, p. 8)	CAP
Khan et al.	2013	Systematic review	Random-effects model	Statin had a beneficial role of reducing the risk of pneumonia.	Incident CAP
Loeb	2010	Systematic review	No statistical synthesis	No direct information from RCTs about the effects of influenza vaccine in preventing community-acquired pneumonia. Pneumococcal vaccine is unlikely to reduce all-cause pneumonia or mortality in immunocompetent adults but may reduce pneumococcal pneumonia in this group.	Incident CAP

Johnstone et al.	2010	Systematic review and meta-analysis	Random effects model	An increased risk of community-acquired pneumonia was found to be associated with PPI use and the duration of PPI use may impact the risk of CAP.	CAP
Engel et al.	2012	Systematic review: Included studies were all prospective cohort studies	Narrative synthesis	CRP alone has limited value to assist the identification of pneumonia cases out of suspected cases but could provide additional diagnostic value based on clinical assessment.	CAP

Table 6.3: Individual predictors identified from systematic review

Study	Older age	Respiratory tract infection symptoms	CRP level/PCT	25(OH)D levels	PPI	Statin	ICS/ Oral steroids	Anticholinergics	β-blockers	CCBs	ACE inhibitors	thiazide	Antibiotic	Anticholinergic	Antipsychotic	Diabetic drugs	Immunosuppressant drugs
(van Vugt et al., 2013)		Yes	Yes														
(Quraishi et al., 2013)				Yes													
(Othman et al., 2016)					No												
(Vinogradova et al., 2011)						Yes											
(Almirall et al., 2010)							Yes	Yes									
(Almirall et al., 2014)																	
(Mukamal et al., 2010)									Yes	Yes	Yes	Yes					
(Jackson et al., 2009)																	
(McKeever et al., 2013)							Yes										
(Paul et al., 2015)														Yes			

(Vila-Corcoles et al., 2012)																		
(Trifiro et al., 2010)																		Yes
(Gessner et al., 2019)																		
(Mullerova et al., 2012)	Yes																	
(Baik et al., 2000)																		
(Steurer et al., 2011)		Yes	Yes															
(Neuman et al., 2010)																		
(Groeneveld et al., 2019)		Yes	Yes															
(Yeh et al., 2019)																		
(Moore et al., 2019)		Yes																
(Kolditz et al., 2019)																		
(Williams et al., 2017)	Yes						Yes											
(McDonald et al., 2015)																		
(Rivero-Calle et al., 2016)	Yes																	
(Lin et al., 2013)	Yes						Yes											
(Evertsen et al., 2010)		Yes																

(Rivero-Calle et al., 2019)																		
(Gorricho et al., 2017)																	Yes	
(Zirk-Sadowski et al., 2018)					Yes													
(Janson et al., 2018)							Yes											
(Marchello et al., 2019)		Yes																
(Baskaran et al., 2019)	Yes																	
(Zhou et al., 2019)				Yes														
(Htun et al., 2019)		Yes	Yes															
(Baskaran et al., 2018)																		
(Walters et al., 2017)																		
(Htar et al., 2017)																		
(Almirall et al., 2017)	Yes				Yes		Yes											Yes
(Schiffner-Rohe et al., 2016)																		
(Lambert et al., 2015)					Yes													
(Abramowitz et al., 2015)					Yes													
(Phung et al., 2013)																		

(Khan et al., 2013)					Yes												
(Loeb, 2010)													Yes				
(Johnstone et al., 2010)					Yes												
(Engel et al., 2012)		Yes	No														
Predictor	6	8	4	2	6	1	6	1	1	1	1	1	1	1	1	1	1
Non-predictor	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6.4: Individual predictors identified from systematic review (continued)

Study	Physical activity	poor functional status	low weight or recent weight loss	excessive weight gain/obesity	Smoking/ Environmental exposures	Dental Hygiene	Winter	Flu vaccination	PPV23	PCV13	cardiopulmonary disease	Co-morbidity	Renal function impairment	Lung cancer	HIV	Previous CAP
(van Vugt et al., 2013)																
(Quraishi et al., 2013)																
(Othman et al., 2016)																
(Vinogradova et al., 2011)																
(Almirall et al., 2010)																
(Almirall et al., 2014)					Yes											
(Mukamal et al., 2010)																
(Jackson et al., 2009)		Yes	Yes								Yes	Yes				
(McKeever et al., 2013)																
(Paul et al., 2015)																

(Vila-Corcoles et al., 2012)									No							
(Trifiro et al., 2010)																
(Gessner et al., 2019)									No							
(Mullerova et al., 2012)												Yes				
(Baik et al., 2000)				Yes	Yes											
(Steurer et al., 2011)	Yes															
(Neuman et al., 2010)	Yes															
(Groeneveld et al., 2019)																
(Yeh et al., 2019)												Yes				
(Moore et al., 2019)																
(Kolditz et al., 2019)										Yes						
(Williams et al., 2017)			Yes				Yes				Yes	Yes				
(McDonald et al., 2015)													Yes			
(Rivero-Calle et al., 2016)												Yes				
(Lin et al., 2013)			Yes								Yes			Yes		
(Evertsen et al., 2010)																

(Rivero-Calle et al., 2019)					Yes	Yes					Yes				Yes	
(Gorricho et al., 2017)																
(Zirk-Sadowski et al., 2018)																
(Janson et al., 2018)											Yes					
(Marchello et al., 2019)																
(Baskaran et al., 2019)					Yes											
(Zhou et al., 2019)																
(Htun et al., 2019)																
(Baskaran et al., 2018)					Yes											
(Walters et al., 2017)										Yes						
(Htar et al., 2017)									Yes							
(Almirall et al., 2017)			Yes		Yes	Yes					Yes					Yes
(Schiffner-Rohe et al., 2016)									No							
(Lambert et al., 2015)																
(Abramowitz et al., 2015)																
(Phung et al., 2013)			Yes	No												

(Khan et al., 2013)																
(Loeb, 2010)								No	No	No						
(Johnstone et al., 2010)																
(Engel et al., 2012)																
Predictor	2	1	4	1	6	2	1	0	1	2	7	4	1	1	1	1
Non-predictor	0	0	0	1	0	0	0	1	3	2	0	0	0	0	0	0

Table 6.5: Results of Quality in Prognosis Studies (QUIPS) assessment for prognostic factors studies for CAP

Author	Year	Sample size	Mean/ Median age	Overall risk of bias	Bias Domains					
					Study Participation	Study Attrition	Prognostic Factor Measurement	Outcome Measurement	Study Confounding	Statistical Analysis and Reporting
Yeh et al.	2019	56,726	49	High	Low	Low	Low	Low	Low	High
van Vugt et al.	2013	2,820	50	High	High	Low	Moderate	Low	Moderate	Low
Quraishi et al.	2013	16,975	43	High	Moderate	Moderate	High	High	Moderate	Low
Othman et al.	2016	320,000	56	Moderate	Low	Low	Moderate	Low	Moderate	Low
Vinogradova et al.	2011	98,239	74	Low	Low	Low	Low	Moderate	Moderate	Low
Almirall et al.	2010	2,662	60	High	Low	Low	Low	Low	Moderate	High
Almirall et al.	2014	1,003	65	High	Low	Low	Moderate	Low	Moderate	High
Mukamal et al.	2010	14,786	58	High	High	Low	Low	Low	Moderate	Low
Jackson et al.	2009	3,519	NA	High	Moderate	Low	Low	Low	High	Low
McKeever et al.	2013	43,169	55.5	Low	Low	Low	Low	Low	Low	Low
Paul et al.	2015	3,061	77	Low	Moderate	Low	Low	Low	Low	Low

Vila-Corcoles et al.	2012	288	73	High	Moderate	Low	Low	Low	High	Moderate
Trifiro et al.	2010	1,947	83	Low	Low	Low	Low	Low	Low	Low
Mullerova et al.	2012	40,414	CAP: 75.1 (10.6); Non- CAP: 70.9 (10.8)	High	Low	Low	Low	Moderate	High	Moderate
Baik et al.	2000	104,491	Men: CAP 61; Non- CAP 56. Women: CAP 37; Non- CAP 36.	High	High	Moderate	Moderate	Low	Low	Moderate
Gessner et al.	2019	84,496	73	Low	Low	Low	Low	Low	Low	Low
Steurer et al.	2011	621	47	Moderate	Moderate	Moderate	Low	Low	Moderate	Low
Neuman et al.	2010	83,165	NA	High	High	Low	Moderate	Low	Moderate	Low
Groeneveld et al.	2019	249	56	High	High	Low	Moderate	Low	Moderate	Moderate

Moore et al.	2019	28,883	37.8% of total cohort were 60 and above	Moderate	Low	Low	Moderate	Low	Moderate	Moderate
Kolditz et al.	2019	45,580	75	Low	Low	Low	Moderate	Low	Low	Moderate
Williams et al.	2017	14,513	70	Moderate	Low	Moderate	Low	Low	Moderate	Moderate
McDonald et al.	2015	191,709	71	Low	Low	Low	Low	Low	Low	Low
Rivero-Calle et al.	2016	2,332,622	61	High	Low	Low	Low	Low	Moderate	High
Lin et al.	2013	2,630	CAP cases: 77.3; Non- CAP cases: 70.1	Low	Low	Low	Low	Low	Low	Moderate
Evertsen et al.	2010	57,704	CAP: 46; bronchiti	High	Low	Low	Moderate	Low	High	Moderate

			s:40; URI: 40							
Rivero-Calle et al.	2019	153,511	61	Low	Low	Low	Low	Low	Low	Low
Gorricho et al.	2017	19,789	72	Low	Low	Moderate	Low	Low	Low	Low
Zirk-Sadowski et al.	2018	150,100	71	Low	Low	Low	Moderate	Low	Moderate	Low
Janson et al.	2018	55,189	66	Low	Low	Moderate	Low	Low	Low	Moderate

Table 6.6: Prognostic effects of predictors identified from low bias primary studies

Author/ Year	Sample size	Study population	Statistic measures	Prognostic factors with effect estimates (95% CI)
(Gessner et al., 2019)	84,496	Elder adult (65 and above)	VPDI (vaccine preventable disease incidence/ 100,000 person-years of observation (PYOs))	PCV13: 72.2 (-5.3 to 149.6)
(Kolditz et al., 2019)	45,580	Patients 60-99 years old	Absolute risk reduction and number need to vaccine (NNV)	PCV-13: ARR 0.63 (0.07 to 1.2; p=0.028); NNV 159 (84 to 1,429)
(Rivero-Calle et al., 2019)	153,511	Adults (18 and above)	Odds ratios (adjusted)	HIV: 5.21 (4.35 to 6.27); COPD: 2.97 (2.84 to 3.12); Asthma: 2.16 (2.07,2.26); Smoking: 1.96 (1.91 to 2.02); Poor dental hygiene: 1.45 (1.41 to 1.49)
(Zirk-Sadowski et al., 2018)	150,100	Adults (60 and above)	Prior event rate ratio (PERR) adjusted net hazard ratio	Proton-Pump Inhibitors (PPI): 1.85 (1.27 to 2.54)
(Janson et al., 2018)	55,189	Adults (40 and above)	Hazard ratio (adjusted)	COPD (FEV ₁ <50% vs FEV ₁ ≥ 50%): 1.33 (1.21 to 1.47); Asthma: 1.13 (1.01 to 1.27);

				ICS use: Low ICS: 1.23 (1.10 to 1.38); High ICS: 1.41 (1.23 to 1.62)
(Gorricho et al., 2017)	19,789	Type-2 diabetes (T2DM) patients	Odds ratio (adjusted)	Oral antidiabetic agents (OADs) in combination with thiazolidinedione use in combination vs metformin plus sulfonylureas: 2.00 (1.22 to 3.28)
(Paul et al., 2015)	3,061	Adults (65 to 94)	Odds ratio (adjusted)	Acute use of anticholinergic medication: 2.55 (2.08 to 3.13); Chronic use of anticholinergic medication: 2.07 (1.68 to 2.54);
(McDonald et al., 2015)	191,709	Adults (65 and above) with diabetes mellitus and no history of renal replacement therapy	IRR (adjusted)	In comparison with eGFRs \geq 60mL/min/1.73m ² , eGFRs<15: 3.04 (2.42 to 3.83), eGFRs 15 to 29: 1.73 (1.57 to 1.92), eGFRs 30 to 44: 1.19 (1.11 to 1.28), eGFRs 45 to 59: 0.95 (0.89 to 1.01); Proteinuria: 1.26 (1.19 to 1.33)
(McKeever et al., 2013)	43,169	Adults (18 to 80) asthma patients	Odds ratio (adjusted)	ICS (\geq 1,000 μ g): 2.04 (1.59 to 2.64)
(Lin et al., 2013)	2,630	Newly diagnosed COPD patients without asthma	Hazard ratio (adjusted)	Age: 1.03 (1.02 to 1.04); Lung cancer: 3.81 (2.88-5.05); Bronchiectasis: 2.46 (1.70-3.55);

				ICS: 1.60 (1.30-1.96).
(Vinogradova et al., 2011)	98,239	Adults (45 and above)	Odds ratio (adjusted)	Statin use in previous year: 0.78 (0.74 to 0.83); Recent statin use (with in 28 days): 0.68 (0.63 to 0.73)
(Trifiro et al., 2010)	1,947	Adults (65 and above) who used an antipsychotic drug	Odds ratio (adjusted)	Atypical antipsychotic (current vs past): 2.61 (1.48 to 4.61); Typical antipsychotic (current vs past): 1.76 (1.22 to 2.53)

6.5 Discussion

In total, there were 33 prognostic factors for incident CAP identified by this study which could broadly be categorized into six groups: **patient characteristics** (older age, respiratory tract infection symptoms), **biomarkers** (CRP, PCT, 25(OH)D levels), **use of medication** (PPI, statin, steroids, anticholinergics, β -blockers, calcium channel blockers (CCBs), ACE inhibitors, thiazide, antibiotics, anticholinergic, antipsychotic, diabetic drugs, immunosuppressant drugs), **life style and environmental exposures** (physical activity, functional status, body weight, smoking/ passive smoking, dental hygiene, season (winter)), **preventive procedures** (flu vaccination, pneumococcal vaccinations (PPV23, PCV13)), **co-morbidity conditions** (cardiopulmonary conditions, renal function impairment, lung cancer, HIV and previous CAP history). Further, identified prognostic factors could be considered as the ones reflecting or affecting host susceptibility, determination of respiratory infectious diseases and possible effective procedures to prevent common causal pathogens.

Apart from CRP, weight gain and obesity, flu and pneumococcal vaccinations, the remaining predictive factors were shown to be conclusive irrespective of the robustness of study quality. For individual prediction research studies as evaluated being low risk of bias, elder age, smoking, poor dental hygiene, the use of PPI, ICS, Oral antidiabetic agents (OADs) in combination with thiazolidinedione, anticholinergic medication, statin and antipsychotic medication, COPD, asthma, bronchiectasis, kidney function impairment, lung cancer and HIV were shown to be candidate predictors for incident CAP. Given that the potential variation in eligibility criteria, prognostic factor measurement and outcome definition, the direction of the prognostic factor effects might be considered more relevant than their effect size. Caution is also required in interpreting the predictive value of variables evaluated by individual studies ranked with high or moderate risk of bias, which might or might not necessarily be considered as invalid. For example, beta-blockers, calcium channel blockers, and lipophilic ACE inhibitors were estimated to contribute to the onset of pneumonia among hypertension patients (Mukamal et al., 2010). This is in line with recent research evidence (Fang et al., 2020, Zheng et al., 2020). The original research was conducted among private insured middle-aged Americans with

hypertension, for which the study population was considered to be highly biased. However, for studies facilitated under healthcare systems where healthcare insurance excises an essential role in healthcare delivery process, such bias is inevitable or should be adjusted during analysis if relevant data is available. Alternatively, quality appraisal could adopt another strategy as recommended by Hayden et al. that assigning overall low risk of bias if an predefined essential domain is rated as low risk of bias (Hayden et al., 2013). But reviewer bias could be introduced by following such recommendation. Therefore, necessary discretion is needed to judge such research evidence critically.

There were inconclusive results for certain predictors including CRP, weight gain and obesity, flu and pneumococcal vaccinations. CRP is not widely ordered in primary care to assist the diagnosis of CAP, but more of providing guidance for antibiotic treatment or predicting adverse events for CAP patients in secondary care (Demir, 2014, Walters et al., 2011). There were conflicting results for overweight and obesity (Phung et al., 2013, Baik et al., 2000) as body weight is a composite outcome of many health-related factors i.e., lifestyle, physical activity, nutrition status, exhaustion due to chronic conditions. As a result, association might be expected to vary according to the specific research topic. With regard to flu and pneumococcal vaccination, evaluation of their protective effectiveness against CAP depends on several factors. The preventive effect of flu vaccination especially among vulnerable population is demonstrated by reduce the risk of annual seasonal flu which would trigger LRTIs. Whereas pneumococcal vaccination mainly targets pneumococcal infections with causal pathogens being the most common but not the only pathogens responsible for CAP. That explains why the effectiveness of pneumococcal vaccines in preventing vaccine-type pneumococcal community-acquired pneumonia is certain but not CAP in general (Bonten et al., 2015).

Like most systematic reviews, this study is subject to several limitations. First, this review deployed a narrative synthesis of prognostic effects of identified variables without meta-analysis. Because this systematic review mainly aimed to inform candidate predictor identification for prediction modelling study. Predictors included in the prediction modelling study were subject to the availability of relevant

information documented in CPRD dataset. Also, the effect size estimations are reserved for the regression modelling study. Second, even if efforts to address publication bias were sought by searching unpublished and incomplete studies, neither grey literature nor incomplete studies were found to assist to reduce potential publication bias (Easterbrook et al., 1991). Finally, the risk of bias assessment for individual prognostic studies may not be optimal given that there is no agreed thorough evaluation tool for prediction research systemic review. This may lead to the exclusion of several studies from low bias group if another approach is used otherwise.

6.6 Conclusions and implications for further research

The prognostic value of candidate factors for incident CAP should be considered when it is able to assist the identification or rule out the diagnosis of CAP; or more accurately reflect individual's ill health which exert influence on the susceptibility to CAP; or demonstrate a strong protective effect against the common causal pathogens of CAP in predefined population cohort in specific settings. Future high quality of prognostic studies in CAP are desirable, so that low bias study results could provide more relevant and robust evidence in this field.

Chapter Seven : Clinical prediction model development for adult RTI patients who reconsulted with pneumonia in 30 days (Introduction and methodology)

This chapter presents the introduction and methodology sections for prediction modelling studies of the thesis with results, discussion and conclusion reported in the following chapter.

7.1 Introduction

Prediction in medicine

People usually makes decisions based on information and experience. Clinical diagnostic practice mainly involves clinical assessment and diagnostic investigation with the latter being employed to confirm clinically suspected diagnoses or to acquire more accurate information (WHO, 2005). The aim of acquiring a concrete diagnosis is to inform further medical intervention and treatment, if available but, even if no treatment is available. Both doctors and patients still want to know the general course of the disease and to tailor such information to their particular situations as much as possible. Sometimes an undesirable health outcome is very unlikely to be altered by medical intervention, but people continue to want to know what to expect and where they are likely to fall on the continuum of possible outcomes, so that well informed decisions could be made. Foreseeing the patient journey after the onset of disease was set to be a cornerstone of medical practice as described by Hippocrates in *'On airs, waters and places'* (Hippocrates and Adams, 1939). In medicine, prognosis mainly refers to the estimation of probabilities or risks of future health outcomes over a predefined range of time among patients sharing similar health conditions. The concept of prognosis can also be applied to the general population through the construction of life tables (Moons et al., 2009). Given the variability among study populations, prognosis is shaped by individuals' clinical and non-clinical features which generally referred to as predictors in prognostic research.

Since a single predictor rarely offers sufficient information for an accurate prognosis, prognostic research generally adopts multivariable approaches (Moons et al., 2009, Steyerberg, 2008, Hemingway et al., 2013, Riley et al., 2013, Steyerberg et al., 2013, Royston et al., 2009). Then probability estimation or risk scoring is based on the combined information from predictor values observed or measured from each individual participant.

When the predictors incorporated into prediction models are associated with the outcome of interest, prediction modelling is by no means designed to generate causal inference, rather prognostic research aims to identify the optimal set of surrogates to explain the outcome of interest. A health outcome depends on the interactions between patients, their social and environmental context and the healthcare system. The prognostic information conveyed by each predictor should be interpreted in the specific context. The variables included in a final prognostic model may not always offer the most accurate classification in comparison with pathology results which provide a valid criterion and reference method to confirm disease status. For example, ethnicity sometimes is adopted as a surrogate of genetic inheritance in some disease studies, whereas in other research ethnicity may be used to indicate certain health behaviors that certain cultures encourage or forbid (Scambler, 2008). Also, factors associated with elevated disease risk may not always lead to a worse prognosis. For example, well-established risk factors for coronary heart disease (CHD) (hypertension, smoking, dyslipidemia, diabetes, and family history of coronary heart disease) may have a reversed relationship to in-hospital mortality among incident acute myocardial infarction (MI) patients without past CVD (Canto et al., 2011). Thus, factors potentially associated with the outcome, without necessarily being causal, could be considered as candidate predictors during prognostic modelling studies. The average effect of each predictor on the outcome of interest, as expressed by the coefficients of prediction model or risk score, could be interpreted as different weights that might be assigned to various patient's profiles during clinical discretion rather the strength of the association which is usually measured by correlation between predictors and outcomes.

The usefulness of a prediction model is a consideration which usually determines the major objective of a prognostic study and the selection of modelling methodology. Certainly, as defined by the fundamental concept of prognosis, prediction models are mainly used to inform doctors and patients during decision making about the future course of possible health events or the probability of developing certain complications. Beyond that, prediction models may be applied in various settings and for many other purposes. By picking out the factors that could explain the onset of the conditions or health status, prediction models depict what kind of patients are most likely to develop the outcome of interest in given period of time. Such information could serve for generating research hypotheses for aetiological or therapeutic studies. By quantifying the probabilities of individual patients, in terms of future events, prediction models enable the stratification of patients at various risk levels informing the design of research studies. Key elements of widely adopted prediction models also help clinicians to have a better understanding about the natural course of ill health. For example, the Glasgow Coma Scale (GCS) is an essential component of Acute Physiology and Chronic Health Evaluation (APACHE) II score for predicting severity of ill health in intensive care unit (ICU) patients, but it has been criticized as having limited utility value in the ICU as it requires the assessment of verbal response, which is absent for many critically ill patients (Dong and Cremer, 2011). The clinical usefulness of GCS index, however, may go beyond whether it is actually being fully applied to every single case, but lies in the concept that brain trauma suggests a worse prognosis for ICU patients and central nervous system function would affect the quality of life profoundly for surviving patients. Bearing in mind of the importance of central nervous system (CNS) functions as reflected by the APACHE II score, ICU medical professionals will be guided to pay attention to neurological signs during daily practice especially when the primary reason for ICU admission is not a neurological condition. Prediction models are also employed for comparing quality of healthcare delivery in the sense that standardized measurements could be made based on that, which enables comparisons to be feasible when direct comparisons are not readily available. Such as the American Society of Anesthesiologists (ASA) score that was primarily used for the classification of physical status of pre-operative patients (Daabiss, 2011), it is also adopted for the surveillance of surgical site infection (SSI) to facilitate the

comparison between patients with different severities undergone same surgical procedures (WHO, 2018a, PHE, 2019).

Machine learning in medical research

As denoted by the meaning of prediction, prognosis is made based on information from a group of people sharing something in common and followed up in predefined meaningful period of time to see what happens to them. This makes the cohort study a suitable study design to address most prognostic questions (Fletcher et al., 2012, Moons et al., 2009). The ideal cohort would be assembled in the present under a well-constructed sampling framework and followed into the future with satisfactory sample size to ensure precision and with only limited attrition to reduce bias. However, it may not always be feasible or possible to conduct research among such cohorts in real life. Given the high cost of prospective cohort studies, 'big data' sourced from EHRs with its wide capture of healthcare information and high volume of data has been brought to the spotlight of medical research (Chen and Asch, 2017). The rudimentary understanding about big data mainly refers to its two notions: the large number of possible relevant variables, the substantial sample size of the data or a combination of both. In addition, 'big data' may also have a greater degree of complexity than traditional epidemiological datasets. As a result, methodologies generally referred to as 'machine learning' methods, which boast the capability to process and analyze the data of high dimension have been brought under the spotlight of medical research deploying big data. Inevitably, machine learning algorithms which are not new in medical research but only empowered by recent advancement of computer performance have been widely implemented to deal with the complexity and the large dimension of 'real-world' data (Efron and Hastie, 2016b). Owing to certain advantages in comparison with most conventional statistical prediction modelling approaches by loosening statistical assumptions, dealing with multiple interactions between predictors and high efficiency to perform analysis on big data, the preeminent component underpinning the excitement of machine learning is its potential to offer information which is less likely distilled by individual clinical wisdom to assist clinical practice and eventually lead to better health (Collins and Moons, 2019). Even though big data and machine learning are

not familiar to most clinicians at first, they are actually natural extensions of conventional statistical approaches that may be familiar to them (Beam and Kohane, 2018).

Machine learning can be considered, an umbrella term for all data-driven approaches across a continuum of decreasing extent of human input. That is, there is no clear-cut distinction between a fully human-specified algorithm, with its analytical behavior completely predetermined, and a machine-dominated model with deep learning being the pre-eminent example, generating models from raw data directly with much less human guidance. Prediction models generated from machine learning algorithms range from disease risk scores such as Framingham cardiovascular risk score to deep learning algorithms outperforming ophthalmologists in detecting diabetic retinopathy (Attia et al., 2019). Such success does not necessarily mean that machine learning is superior; rather, empowered by enormous data volumes and high-performance computational resources, machine learning algorithms may risk compromising interpretability and transparency of the outputs. High-volume data streams may make it possible to capture the complexity and, heterogeneity of the raw data and dilute the impact of non-informative variables especially when it comes to ‘real-world’ data. Also, the specification of models relies on the performance of computational resources due to the nature of machine learning processes, particularly for deep learning. Just because the major task of model development has been shouldered by the algorithms, how the algorithms function resembles a ‘black box’ to human intelligence, and critically appraising the outputs would be challenging to most intended users (Watson et al., 2019).

Apart from the lack of interpretability, machine learning has been criticized for its increasing implementation in healthcare research. One major controversial concern is its marginal benefit compared to traditional statistical modelling approaches including simple linear and logistic regression models. However, most comparative studies drew on well-structured data, including laboratory data or clinical trial data, to develop prediction models (Horne et al., 2009). It is not surprising that machine learning techniques may not perform substantially better than traditional statistical approaches since the predictive information conveyed by pre-defined variables was

informative enough to answer research question irrespective whether the model was derived using a simple logistic regression model or a novel machine learning algorithm. Additionally, eligible predictors were often narrowed down into a restricted scope based on prior knowledge about biomedical mechanisms, or even through use of pilot studies. Therefore, the strengths of machine learning were not demonstrable in this scenario as machine learning is able to deal with less structured, 'messy' real-world data to make the 'unseen pattern' visible.

Given machine learning methods are commonly applied to existing data, questions concerning data aggregation begin to arise (Chen and Asch, 2017). Continuing innovation of medical science and technology means that, medical practice is always changing. Study results have to be viewed in a dynamic social context, so that medical research is chasing an evolving target (Chen and Asch, 2017). That is, if the past may not necessarily resemble the future, then how many years' data should be compiled together for model development, or which study design could eliminate or adjust for such effects on the prognostic model? Given the statement of *'Everything changes but change itself'* (Hussey, 1999), chronological change and questions of validity arising from it are unavoidable in prognostic research studies (Chen et al., 2017). The option basically lies in the trade-off between sample size and underlying effect of dated information. Google flu provides a real-life example to illustrate that using accurate latest point data works better than using accumulated historical data to predict the future (Lazer et al., 2014). Such effects of chronological change also link to the time span of predicted future as there may be meaningful changes anticipated in future. That is, a prediction model usually works better for outcomes in the short term than those in distant future (Chen and Asch, 2017, Moons et al., 2014). The time span of the future period varies from hours (i.e. prediction of post-operative complications) (Canet et al., 2010) to years, as in the classic Framingham Risk Score for 10-year CVD risk estimation (Kannel et al., 1976). For prognostic studies among post-operative patients, 48 or 24 hours or even further level of granularity might be defined to help perioperative patient care delivery, place of treatment (general surgery ward, post anesthesia recovery unit (PACU) or surgical intensive care unit (SICU)), improving the allocation of scarce healthcare resources (Arozullah et al., 2000). Whereas for palliative care, weeks, months or years are generally the frame of

reference for planning the last stage of life comparing to the exact date of eventual death- the preciseness of which is rarely necessary. The effect of time or time interval on the development of predefined outcome should also be taken into consideration during methodology selection. If hazardous effects of predictors remain constant over time, survival analysis and a time-to-event framework will often be appropriate for the proposed research question (George et al., 2014); if the nature of disease progression is rapid i.e. severe sepsis in ICU patients, where current health status is closely linked to the most recent data, then Markov models may be more likely to yield optimal decision supporting tools using real-time monitoring data (Komorowski et al., 2018). Being aware of such potential impacts will contribute to framing prognostic research questions, selecting modelling methodology, interpreting prognostic study results, and conducting model validations especially external validation, updating existing model and possible application scope of developed models.

Machine learning could also be interpreted as using a group of computer algorithms to quantitatively explore suitable information at hand to learn from data. The process of gathering patient medical history, physical examination, clinical tests, laboratory experiments could be regarded as primary data collection from a research point of view. Apart from methodological innovation, the performance of machine learning results is also inherently determined by the characteristics and quality of the data being 'fed' to the algorithms. Even if machine learning is a powerful tool for many research questions, it is not a panacea for any type of data-driven research study nor able to squeeze out information that is not documented in the dataset. Additionally, the interpretation of machine learning outputs should be put into the wider context relevant to the information documented in the data set. Given that the majority of data machine learning techniques being applied were not developed for medical research purposes, data exploration should be done prior to handing on the data for model development, in order to address the quality and bias of data. Using clear case definitions is a necessary approach to make the data analyzable and in return to have clear answers to research questions but understanding the underlying concept behind the stringent definitions are equally important.

Concerns about machine learning model evaluation have drawn attention to the calibration of performance as this may not be well investigated or reported transparently (Shah et al., 2018). This is not unique to machine learning but generic to all prediction models. Prediction model performance mainly comprises two essential elements: discrimination performance, as assessed by concordance statistics including c statistic or its equivalent of AUROC, and calibration performance quantified by H-L test or goodness-of-fit tests (Steyerberg et al., 2010).

Discrimination performance describes how well the model picks out the right individuals from their parent cohort that is cases from the true case group or non-cases from the non-case family. This performance is more related to identified variables included in the final model, whereas for calibration performance assesses the alignment between predicted outcomes and the observed events, which may be heavily influenced by unknown variables not included in the model. Calibration contributes to reliability in clinical practice for decisions derived from prediction algorithms, which could be more essential than precision. If the unincluded variables contribute more when a prediction model is applied to another population, then the prediction model will be lack of clinical usefulness and even introduce the danger of potential harm to patients (Shah et al., 2018).

In conclusion, the questions and concerns raised in the field of machine learning and big data actually represent generic fundamental considerations in research including relevant research question, suitable research material and feasible research methodology to process the raw data into a sensible answer and requiring caution in the application of research output. Machine learning comprises a group of analytical tools, how the techniques are used should be an important consideration during research design phase. Understanding the nature of healthcare information being used will help to interpret the underlying meaning of certain variables represented given the context of the study dataset and overarching conditions such as the healthcare system, influential health policies. Use of the prediction model should be built upon the understanding of research question, research data and methodology. Setting up a realistic goal of machine learning approach is helpful to stay away from the inflated expectations in this big data era.

General considerations for prediction model development

During prognostic model development, researchers initially have to identify candidate variables for the final model. The generic rule for variable selection is to take both statistical and clinical significances into consideration with the aim of translating research results into better healthcare. Statistical methodologies adopted during prediction modelling mainly serve for two purposes: To answer the research questions and quantify the accuracy of estimates. These processes will allow sensible inferences to be made and will eventually gauge the clinical usefulness of final model. Clinical epidemiological studies are always accompanied by variances and biases. By characterizing the accuracy of the estimates and quantifying whether bias is sufficiently large to result in an alteration of clinical significance, clinicians or other intended users will be assisted in deciding how much to rely on the model during decision-making for each individual patient. Similarly, predictor selection should not only use statistical importance as reference because included predictors must offer clinically relevant information.

Another general consideration for prediction model development is how to deal with over-fitting. The final prediction model aims to make predictions for out-of-sample populations. That is, the usefulness of the model will depend on how well it works after being applied to those whose information was not used for model development. If a prediction model performs perfectly on the sample from which it derived, it may not necessarily perform well on unseen out-of-sample data. The extreme over-fitting case would be that the model is able to describe each individual by each combination of the predictors included, just like a finely tailored handmade suit which is quite unlikely to fit well on people other than its owner. In order to address this problem, several methods have been developed. First, simple models generally work better on out-of-sample than complex models. Therefore, machine learning techniques are able to penalize model complexity known as ‘regulation’ in regression-based models or ‘tuning’ for tree-based approaches, so as to reduce excessive complexity. Second, since over-fitting relates to ‘unseen’ samples, a conventional approach would divide the data into training, validation and testing sets. Training data is used for model estimation, validation data for model selection, testing data for apparent model

performance evaluation. Sometimes, validation and testing sets could be combined. In the context of big data, where data is sufficient enough, cross-validation (partition the data into training and validation sets) makes more sense than train-test exercise as the former emphasizes the out-of-sample performance rather than in-sample performance as reflected by the latter one. Third, if a penalized or tuning parameter is needed for the optimal out-of-sample prediction, k-fold cross-validation is proposed to identify the value of this parameter. Usually, 5 or 10 are recommended for k and the maximum number is the total sample size minus one ('leave one out'). Then a series of values of the penalty or tuning parameter and its correspondence loss (as measured by mean squared error (MSE) or misclassification error) are generated, which enables the selection of appropriate value for the parameter (discussed in methodology for penalized regression later) (Varian, 2014). Further, ensemble methods like bootstrap, boosting and bagging are also adopted to deal with over-fitting problems by adding 'randomness' to the model as illustrated in Table 7.1. Sometimes, researchers ask a question that may be quite difficult for the resource available to answer irrespective if machine learning is employed or not.

Table 7.1: Illustration of common ensemble methods

Ensemble methods
Bootstrap randomly samples a number of subsets from the whole population with replacement repetitively at sufficient large of times. Then out-of-sample estimation is made for certain statistic distribution.
Bagging is short for bootstrapping and aggregating which denotes the algorithm works out multiple models based on bootstrap samples but averages or takes the majority votes across the bootstrap sub-sample models.
Boosting is a technique which convert a 'weak' learner into a 'strong' one by assigning increasing weight to misclassified observations during iterative repetition. The common algorithms from this family are Gradient Boosting (GBM) and XGboost (regularized boosting) for tree-based approaches. Unlike bagging, the models estimated based on bootstrap sub-samples are not independent from each other as continuous corrections are made on top of the previous models.

Supervised machine learning approach to predict RTI patients who reconsult with pneumonia in 30 days

For this prognostic study, clinical prediction modelling can be regarded as a process of using healthcare information from electronic health records to characterize patients presenting to the GP practice with evidence of RTIs who are likely to reconsult with pneumonia within 30 days. The prediction modelling process of this thesis followed the seven steps as recommended by Steyerberg and Vergouwe (Steyerberg and Vergouwe, 2014), constructed according to the recommendations of the Prognosis research strategy (PROGRESS) framework (Hemingway et al., 2013) and reported in comply with the TRIPOD statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (Collins et al., 2015) (Table A 33) alongside with RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) guidelines for EHR studies (Benchimol et al., 2015). Key questions involved in each modelling step (Steyerberg and Vergouwe, 2014) for this thesis are addressed below:

Step 1: Research question definition and data inspection

- a. What is the precise research question?

Given that this prognostic study is not an aetiological study to establish causal inference, the term ‘re-consult’ was used instead of ‘develop’ even though pneumonia was the most common infectious complication after RTIs (Gulliford et al., 2016, Tan et al., 2008). To explore whether pneumonia resulted directly from treatment failure of preceding RTI events or early onset of other diseases, for example pulmonary cancer, is not the major interest of the developed model. Instead, quantitatively describing subgroups of RTI patients who are at high risk of being diagnosed with pneumonia within 30 days will be the focus. This will inform general practitioners of patients for whom timely treatment, upgraded treatment plan or clinical tests should be offered so that undesirable health events could be avoided, or early detection of other treatable conditions is able to be facilitated. The time-interval of 30

days was chosen according to the natural course of disease after factoring in the average waiting time of GP appointments in the UK primary care system (Tan et al., 2008, Primary Care Domain and Service, 2018). Additionally, the major interest of the prediction model is to predict the onset of pneumonia after RTIs rather than the multiple subsequent episodes of pneumonia, therefore the first incident pneumonia event in each calendar year is sought after during individual case identification.

b. Predictor identification

This prognostic study adopted an inclusive approach during initial candidate predictor identification based on the systematic literature review of previous study results, as well as advice from medical professionals. That is, apart from the objective information offered by statistical analysis and clinical importance documented in existing evidence, findings in the data were also interpreted with reference to subjective knowledge of practicing medical professionals within the research group. Any possible relevant information to current research evidence from disease management guidelines was also included. For example, frailty was adopted to describe the overall health status of individual patients as NICE guideline signifies that *'if the patient is systematically very unwell'* is at higher risk of developing infectious complications (Tan et al., 2008, Hanlon et al., 2018). Also, the guideline identified *'current use of oral glucocorticoids'* as another risk factor for pneumonia in older patients. Therefore, all immunosuppressive conditions and drug prescriptions i.e., 'Total splenectomy' and 'Cancer chemotherapy' were incorporated during predictor sorting to reflect the immunosuppressive statement of individual patients.

c. Case ascertainment

Clinically diagnosed pneumonia was adopted as the primary endpoint during case selection. Even though the previous study suggested that clinically suspected pneumonia as defined by chest infection treated with antibiotics

was more prevalent than pneumonia in the UK primary care settings and its essential elements during clinical disease management are similar to those of pneumonia, chest infection was investigated initially as a predictor following Petersen et al. (Petersen et al., 2007a). This is because chest infection comprises two major clinical scenarios: pneumonia and bronchitis with the later one generally considered to be self-limiting, and there are no concrete recommendations from any academic bodies suggesting that a chest infection diagnosis is equivalent to pneumonia. Previous studies have demonstrated that pneumonia is the most frequent infectious respiratory complications after chest infection. Subgroup analysis was conducted to find out if there were systematic differences between chest infection patients and other self-limiting RTI patients in terms of the future probabilities of pneumonia consultation within 30 days.

Another consideration during case ascertainment was given to the frequency of endpoint which is essential to determine the effective sample size rather than total sample size as denoted by events per variable (EPV). When outcome is binary, EPV is calculated as the smaller of the number of study population who experienced and who did not experience the outcome interested divided by the number of predictors considered to be included in model development. The value of EPV is recommended varying from 10 to 20, even suggested to be relaxed below 10 when confidence interval coverage and bias are within acceptable level (Austin and Steyerberg, 2017, Hickey et al., 2018, Vittinghoff and McCulloch, 2006). In the thesis, there were 36 candidate variables for model development, therefore 720 events were needed if EPV of 20 is adopted. Given that there were 16,289 pneumonia re-consultation cases included for prediction modelling study, this criterion was more than adequately met.

Recently, the blanket application of EVP has been criticized for being too simplistic and other criteria of minimum sample size for prediction model development have been proposed (Riley et al., 2019b). Riley et al. have recommended three criteria when calculating sample size for logistic or the

cox model development: 1.) a global shrinkage factor no less than 0.9; 2.) the maximum difference between apparent and adjusted Nagelkerke's R^2 is suggested to be 0.5; 3.) preciseness of overall risk estimation in the population (Riley et al., 2019b). According to their recommendations, a minimum sample size of 3,056 is needed for this prediction modelling study. Therefore, the sample size of our prediction modelling study was considered to be ample.

d. Addressing treatment effects

Predictors with potential treatment effect could vary from single predictor to combination of predictors with responsiveness to relevant interventions (Attia et al., 2019). Treatment here may not only refer to specific medical treatments, rather any type of procedure or characteristics that would modify individual's probability of benefiting or experiencing harm from a specific treatment. Predictors of treatment effect should be investigated typically for prediction models aiming to predict treatment effects of particular intervention so that risk stratified clinical decision could be applied to various groups of patients, then treatment could be offered to those who most likely benefit from it or held back from patients for whom potential harms will offset the clinical improvements (Hingorani et al., 2013, Riley et al., 2013). Predictors of treatment effect might differentiate patients who experienced meaningful positive or negative effects between offering and withholding the treatment or between various alternative treatments. For certain predictors, they could be both predictor for the disease and the predictor of treatment effect for the disease. For example, age greatly influence the risk of stroke onset among atrial fibrillation patient without anticoagulation treatment. Meanwhile, age also predicts patient's response to anticoagulants (Zhao et al., 2016). On the other hand, certain factors could be considered as predictor only at the presence of treatment. For example, the route and timing of administration for prophylactic use of antibiotics to evaluate surgical site infection prevention effectiveness (Hawn et al., 2013, Darouiche et al., 2012). Generally, predictors of treatment effect reflect the underlying characteristics

of target population in terms of the possible response for specific treatment(s).

However, there is no straightforward solution to identify and test predictors of treatment effect. Several study designs have been proposed to address the issue of predictor of treatment effect including treatment only, test-positive only, or non-randomized comparative study designs (Riley et al., 2019c). For this study, predictors of treatment effect to predict the onset of pneumonia mainly relates to antibiotic prescription during initial RTI consultations. Since the assumption that immediate antibiotic treatment alone would be sufficient to prevent a subsequent pneumonia event may not always be applicable for several reasons including the empirical selection of antibiotics, patient's compliance of treatment. Therefore, RTI consultations were evaluated further according to antibiotic prescription on the index date to explore if antibiotic treated and non-treated patients were at apparent different risk groups in terms of the outcome of interest.

e. Handling missing data

Missing data is an unavoidable issue for prognostic research using EHRs from case ascertainment to predictor identification (Riley et al., 2019c). Especially for 'soft' healthcare information, such as lifestyle preferences, physical activity, health behavior and mental health status even certain measurements such as body mass index (BMI), the data are generally not missing at random. Apart from 'hard' endpoints like mortality data, even the absence of certain disease confirmation information may not be confidently considered as missing values. Since the identification of ill health largely depends on patient's health seeking behavior, if the patient never presented to clinical consultation with certain chief complains, clinical manifestations of pre-existing health conditions were not evident enough to draw the attention of clinicians or patients for early detection. Therefore, healthcare information that was documented during daily practice is prone to be missing selectively.

There is no perfect solution to avoid bias introduced by such missing information (Riley et al., 2019c). Several approaches have been proposed to deal with miss data problem in EHRs research. Complete case analysis tends to give unbiased estimation for outcome endpoints missing completely at random (Groenwold et al., 2011), during which cases with missing data are excluded from the analysis. This is not suitable for this prognostic research as a substantial number of self-limiting RTI patients will be excluded from the study simply because they had better general health status without long-term morbidities being recorded. For this prediction modelling research, missing indicator (Groenwold et al., 2012b) together with crude imputation (Riley et al., 2019c) were performed parallelly, so that comparison was made to explore possible predictive information of missing data within individual variables. Missing values were categorized as a single group in the missing indicator approach, whereas in crude imputation the missing values were treated as ‘normal’ or the average category.

Multiple imputation is generally preferred to address missing data problem by mapping the underlying distribution and predicting missing values using available information (Groenwold et al., 2012b), which is generally preferred to complete case analysis even when data are MNAR (Kontopantelis et al., 2017). However, this method may not be appropriate for this prognostic study. Because the decisions to record certain medical information by clinicians are usually based on medial professional knowledge and previous ill health history as well as patient’s presentation during consultations. Healthcare information that was not recorded in EHR sometimes simply because of the absence of medical conditions, which might not be treated as missing data and carry meaning (Steele et al., 2018). Additionally, the proportion of missing data for certain predictors was often considerably more than 5%. Therefore, multiple imputation was not performed in this thesis because it does not fit for the assumption of the methodology (Sterne et al., 2009a).

Step 2: Categorization, re-coding and re-grouping predictors

The generic rule for handling predictors is to keep continuous data continuous especially during initial data inspection and exploration phase so as to retain as much information as possible (Royston et al., 2006, Altman and Royston, 2006, Altman, 2014). Potential non-linear relationships between continuous variable and outcome of interest were explored using descriptive statistics together with CART modelling. The choice of modelling predictors with non-linear associations mainly lies in between using restricted cubic spline (Harrell Jr, 2015), fractional polynomials (Royston and Sauerbrei, 2008) and categorization according to several cut-off values with statistical or clinical significance (Boersma et al., 2001). For example, the association between age and pneumonia incidence rates is ‘U’ shape in the general population, when the study population was set at adult population, the non-linear relationship is presented as ‘J’ shape suggesting age exerts accelerating risk effects for the onset of pneumonia among subgroups of patients. In the later phase of prediction model development, age was categorized as 16-35, 36-45, 46-55, 56- 65, 66-75, 76-85, 85 and above. This enables the model to be user friendly with essential concepts in line with current disease management recommendations after factoring the results from variable selection.

An iterative approach was adopted during data re-coding and re-grouping. Predictors with overlapping information were kept as separate categories i.e., frailty index score and severity groups based on that were included as individual predictors during initial variable selection process. RTI subgroups were re-coded based on clinical guidelines as well as previous research evidence. For example, cough, chest infection and LRTIs were investigated in various studies especially focusing on antibiotic treatment for respiratory conditions (Petersen et al., 2007a, McDonagh et al., 2016, NICE, 2015, Troeger et al., 2018). Then, RTI conditions were further grouped into broad categories based on preliminary results after referencing medical literature, considering the ease of clinical usefulness or the relevance to answer specific research questions.

Step 3: Model specification

Given incident pneumonia patients are identifiable from the dataset, therefore supervised machine learning approaches were applied to variable selection of modelling process (Kotsiantis et al., 2007), which are described in a later section.

Step 4: Model estimation

After model specification, the average effects of selected predictors on the outcome of interest were quantified by simple logistic regression modelling methods. Both simple logistic regression and CART models were performed.

Step 5: Model performance

Model discrimination performance was assessed using area under the receiver operating curve (ROC curve) and calibration performance was examined by Hosmer-Lemeshow goodness-of-fit test (Vickers and Elkin, 2006). Cross validation strategy was adopted during model performance evaluation.

Step 6: Model validation

Data were randomly split into 80% for training dataset and 20% for test dataset during internal validation. Temporal validation using recent 25% subset was adopted to investigate temporal stability of baseline risk and predictor effects, which reflects the transportability of developed prediction models (Austin et al., 2017).

Step 7: Model presentation

CART was employed as the main graphic format to communicate prognostic research results with intended audience.

7.2 Methodology

7.2.1 Study design

A stratified sampling approach was adopted for prediction modelling development. This allowed prediction modelling to be carried out with sufficient sample size meanwhile to address class imbalance (Kuhn and Johnson, 2013) between the incidence rates of RTI patients who re-consulted with pneumonia within 30 days and those did not. The predictive value of a prediction model is not a property of the model alone which is also influenced by the prevalence of the disease among the study population (Fletcher et al., 2012). Since the prediction model aims to identify individuals who are more likely to re-consult with pneumonia within 30 days among RTI patients, using more information from non-cases with significant higher incidence rates will result in a model with high specificity and low sensitivity, which is less desirable irrespective to the overall model performance. Given the final model serves to aid clinical discretion at the point of care, factors that could not be addressed at individual consultation scenario i.e., regional variation in health care delivery, chronological change of disease management as identified in chapter five were chosen to be stratified sampling criteria, so that their effects on the outcome of interest were not investigated in this project. Therefore, RTI patients free from pneumonia diagnosis within 30 days were randomly selected from the same practice and study year of those presented with pneumonia within 30 days subsequent to initial RTI consultations. The initial sampled population was four times proportionate to pneumonia patients in the same stratum.

7.2.2 Data source

The UK Clinical Practice Datalink (CPRD) was adopted as the main data source for prediction modelling development. CPRD is the largest database of primary care health care records with a coverage of approximately 7% of the total general practices (Williams et al., 2012). About 98% of the UK population are registered with general practice, also universal healthcare system in the UK made most common conditions be managed in the community. Therefore, healthcare

information documented in CPRD could be considered as a representative sample of the UK population healthcare profile (Wolf et al., 2019). Sixteen calendar year data from the beginning of 2002 to the end of 2017 was aggregated for prediction model development and validation. Information documented in free text was not accessible via CPRD since 2013 due to patient privacy and data governance considerations (Wolf, 2018).

7.2.3 Sample selection

Pneumonia cases were defined using Read codes relevant to ‘pneumonia’ terms after excluding influenza, tuberculosis (TB), fungal and parasite pneumonia. The final code list was adopted from chapter five which was reviewed and finalized by two researchers with clinical and epidemiological backgrounds independently. All pneumonia cases documented in the dataset with eligibility for data linkage were identified during the study period as outlined in chapter five. Diagnostic records one year after the starting date that patients started to contribute to CPRD were included to eliminate any possible duplications during patient registration and transient resident records. Then RTI consultation records were traced back up to 30 days before pneumonia index date. Pneumonia cases documented one day after the RTI consultations were also excluded to eliminate artefactual cases due to delayed data entry. Recurrent pneumonia incidence events for the same patient in the same year were differentiated using 90-day window period. Only the first episode was included as recurrent pneumonia onset is not the major interest for this thesis. Finally, study population for adult cohort was confined within patients aged 16 and above as shown in Figure 7.1.

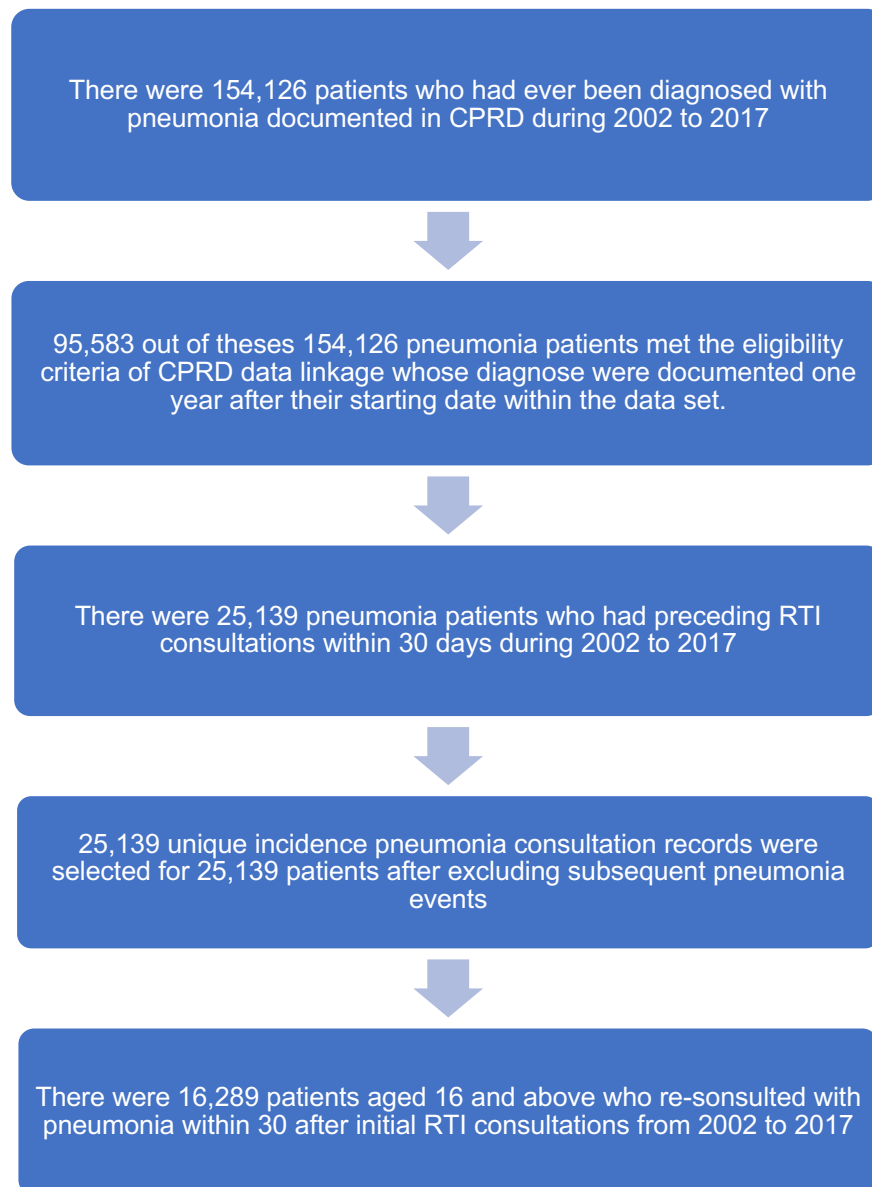


Figure 7.1: Flow chart of pneumonia case ascertainment

In order to identify comparison patients who consulted with RTI but did not develop pneumonia, a define query was initially run in the CPRD database to identify patients who consulted with RTIs during the period 2002 to 2017. There were 5,915,225 patients with at least one RTI consultation. The sample was then restricted to those patients who contributed up-to-standard records during the study years and met data linkage eligibility. Consultation records documented one year after the starting date when patients began to contribute to CPRD were selected.

Then, a random sample was drawn of RTI patients that were eligible for CPRD data linkage of all ages using the ‘sample’ command in R. The sample was stratified by

general practice and index year. There was no stratification by age and gender so that these variables could be evaluated as predictors. Up to four times non-pneumonia patients were randomly selected proportionate to pneumonia cases in each stratum (the 23,490 patients rather than adult cohort only) based on practice ID and study year of pneumonia cases. All RTI consultation records of sampled patients were identified from CPRD then narrowed down to the ones within the study period. Sampled non-pneumonia patients that were also found in pneumonia patient cohort in different study years were excluded to retain as much information as possible from the pneumonia cases. Finally, single one consultation record was randomly selected for each patient as shown in Figure 7.2.

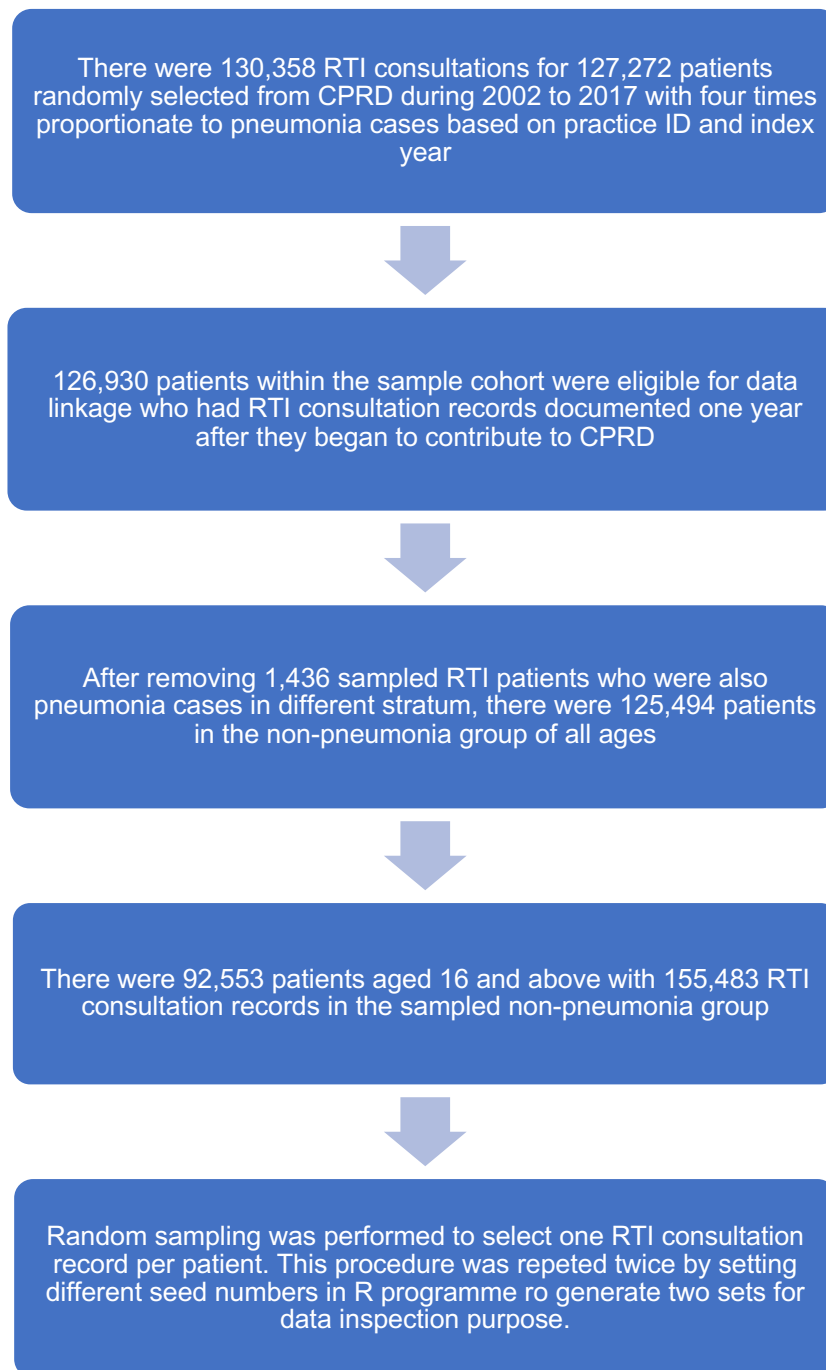


Figure 7.2: Flow charts of non-pneumonia case sample selection based on pre-defined stratified sampling criteria

7.2.4 Predictor definition

An inclusive approach was adopted for candidate predictor specification based on systematic literature review and clinical guidelines. Included predictors were sorted based on the availability of information documented in CPRD data set. During data processing, continuous variables were kept continuously, and for categorical variables derived from any scoring algorithms i.e., frailty category, both the categories and original scores were retained to explore any possible predictive information. Recoding and further data binning were performed using statistical results and clinical implications as reference. The main categories of candidate predictors were clustered following the process of clinical discretion as shown below:

7.2.4.1 Demographic information

Age: given pneumonia disproportionally affect the young and elder population, age exercises an important role in the onset of pneumonia aetiology. Age was calculated as RTI index year minus patient's year of birth. (Clinical and Referral files for RTI index date; Patient file for patients' year of birth)

Sex: according to previous study results, higher incidence rates of clinically diagnosed pneumonia were found among male population, whereas higher clinically suspected pneumonia incidence rates were presented in female cohort. (Patient file)

Season: effects of climate factors on respiratory tract infections have been well documented in existing research evidence (Loh et al., 2011), especially certain causal viral pathogens were detected with winter peaks (Srivastava, 2010, Rossi et al., 2007). Also, weather conditions also relate to outdoor activities including health-care seeking behaviour as well as air moisture affecting the susceptibilities of upper airway mucosa to infection (Eccles, 2002). Meteorological seasons were defined according to UK met office as spring (March, April, May), summer (June, July, August), autumn (September, October, November) and winter (December, January,

February) (MetOffice). (Season was identified by the calendar month of the RTI index date from Clinical and Referral files)

Smoking: smoking has been identified as a well-known hazardous factor of many health conditions particularly for respiratory system (Arcavi and Benowitz, 2004, Fielding, 1985, Yanbaeva et al., 2007). Smoking status was classified as non-smoker, ex-smoker and current smoker. (Smoking status and smoking cessation were identified from Clinical, Referral and Additional Clinical Details files; replacement treatment and smoking cessation information for current smokers such as nicotine replacement and varenicline prescriptions were sought after from Therapy file) Medical and product code lists were adopted from Booth et al (Booth et al., 2013, Booth et al., 2015).

7.2.4.2 General health

Frailty: frailty is recognized an age-related clinical biological syndrome with great heterogeneity in terms of the trajectory from onset and development across several organ systems (Rodriguez-Mañas and Fried, 2015, Dent et al., 2019). It has been characterized as deterioration in physiological capacity (weakness, slowness, physical inactivity and exhaustion) leading to increased vulnerability to stressors (such as respiratory infection in this case) (Clegg et al., 2013, Junius-Walker et al., 2018, Fried et al., 2001, Rockwood and Mitnitski, 2011). Apart from variation in the natural history of frailty, an array of instruments for frailty identification i.e., frailty phenotype, frailty index, Comprehensive Geriatric Assessment (CGA) with different implications as well as possible feasible clinical setting have been proposed (Dent et al., 2019, Hoogendijk et al., 2019, Gulliford, 2019). Even if the concept and utility of frailty have been challenged due to its ambiguous definition, questions around the liability, sensitivity and specificity of its assessment, frailty still is perceived to be meaningful to clinical practice and research (Hanlon et al., 2018, Kojima et al., 2017). In this prediction modelling study, predictive value of frailty was explored while recognising its limitations since it potentially describes patient's ill health using accumulated non-specific clinical manifestations rather than well-established clinical diagnoses like morbidities, which may offer indirect measurement at latent

stage of certain diseases preceding the identification of underlying conditions. Additionally, frailty and pneumonia share similar increasing trends in the adult population with ageing suggesting possible associations (Hoogendijk et al., 2019, Hoogendijk et al., 2016, Sun et al., 2019). During prediction modelling, the electronic frailty index (eFI) developed by Clegg et al (Clegg et al., 2016), which based on 36 clinical deficits together with its code list was adopted due to its successful application in the UK primary care electronic health records including CPRD (Ravindrarajah et al., 2018, Ravindrarajah et al., 2017). Each deficit was identified by medical codes from Clinical and Referral files; eFI was calculated by the count of individual deficits in each study year divided by 36, and patients were classified as fit if $eFI \leq 0.12$, mild frail if $0.12 < eFI \leq 0.24$, moderate frail if $0.24 < eFI \leq 0.36$ and severe frail if $eFI > 0.36$.

Co-morbidity was evaluated as present or absent for individual patients in each study year. Two methods for classifying comorbidity were adopted during data processing. Comorbidity was initially classified using the code list used to identify people as eligible for flu vaccination. The ‘seasonal flu at risk Read Code’ list (NHS and BMA, 2015a) includes diagnostic codes for coronary heart disease (CHD), chronic kidney disease (CKD), chronic liver disease (CLD), chronic neurological disease (CND), chronic respiratory disease (CRD), diabetes mellitus (DM) and disorders of the immune system (Immune condition). Drug product codes were modified as: asthma treatment agents including bronchodilators and inhaled corticosteroids, systematically administrated corticosteroid drugs and immunosuppressive drugs. A second comorbidity classification was adapted from the Charlson Index using a Read code list developed by Khan et al. (Khan et al., 2010). Both the weighted Charlson comorbidity index and 19 categories of comorbid conditions (including myocardial infarction, congestive heart failure, peripheral vascular disease, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease, mild liver disease, diabetes without chronic complication, diabetes with chronic complication, hemiplegia or paraplegia, renal disease, any malignancy without metastasis, leukaemia, lymphoma, moderate or severe liver disease, metastatic solid tumour, AIDS) (Charlson et al., 1987) were identified for each patient. During the model development process, both composite endpoints measured

as comorbidity condition counts and weighted Charlson Index together with individual chronic conditions, medication categories were included at the initial stage to explore which measurement could offer predictive information to the outcome of interest. All comorbidity conditions were identified by medical codes from Clinical and Referral files; medication categories were identified using product codes from Therapy files.

7.2.4.3 RTIs and related respiratory conditions

According to NICE guidelines and previous study results, RTIs and related conditions were initially grouped into eight categories: Cold/ Influenza/ URTI, Sore throat/Pharyngitis/Tonsillitis, Rhinosinusitis, Otitis media, Cough, Bronchitis, LRTI and Chest infection (NICE, 2008b, NICE, 2015). All RTI consultation records were identified from Clinical and Referral files using code list from previous studies (Sun and Gulliford, 2019, Sun et al., 2019).

7.2.4.4 Medical management

Immunization: there were three main type of vaccines covered by NHS vaccination scheme aiming to offer protection against respiratory infectious causal pathogens (NHS, 2019b): Haemophilus influenza type b (Hib), Streptococcus pneumoniae; and flu vaccinations. The seasonal flu vaccination is updated annually, and the vaccine is developed based on the annual assessment indicating the probable prevalent types in the northern hemisphere regions. Only pneumococcal and flu vaccines were given to adult population and especially recommended to those aged 65 and above (NHS, 2019b, NHS, 2019c, NHS, 2019a, WHO, 2019b).

Flu vaccination uptake was defined from 7 days before RTI index date up to one year previously. The flu vaccination code list was adopted from seasonal influenza vaccination programme (NHS and BMA, 2015b) and Leite et al. (Leite et al., 2017). Flu vaccination records were identified from Clinical and Referral files by medical code with administration status being 'Given'; the Therapy file using product codes for vaccine prescriptions; and the Immunisation file using immunisation type.

Pneumococcal vaccinations have shown protective effects against invasive pneumococcal disease and vaccine type CAP (CDC, 2017, Pilishvili and Bennett, 2015, Bonten et al., 2015). Pneumococcal vaccination uptake was defined as any pneumococcal vaccination record documented 10 days before RTI index date. Pneumococcal vaccination codes were searched from CPRD files with flu vaccination records as guided by Leite et al (Leite et al., 2017). Pneumococcal vaccination records were identified from Clinical and Referral files by medical code with administration status being 'Given'; and Therapy file using product codes for vaccine prescriptions; and Immunisation file using immunisation type.

Antibiotic treatment has been shown to have treatment effects for some cases of RTIs and might protect patients from possible severe bacterial infectious complications (Petersen et al., 2007b, NICE, 2008b, NICE, 2015). During prediction modelling, both antibiotic prescription on RTI index date (or next day) and antibiotic prescriptions in the subsequent 30 days for non-pneumonia cases, or before pneumonia re-consultation date for pneumonia cases were included. Apart from any treatment effect, antibiotic prescriptions issued on the RTI index day, together with those documented subsequently reflected clinical discretion and might indicate the clinician's perception of an existing bacterial respiratory infection or a perceived higher risk of developing severe bacterial complications. Antibiotic prescription records were identified from the Therapy files with issued date shown as date of event using product codes from previous study (Sun and Gulliford, 2019). Antibiotic prescriptions issued within the predefined 30 days for any other causes were analysed as a separate category labelled as 'Antibiotic ever'.

Clinical investigations ordered or performed during RTI consultation were included. It was separated into two broad categories: 'clinical check' such as clinical follow-up of chronic conditions including blood pressure check for hypertension patients, and 'clinical test' which included all types of laboratory tests including chest X-ray and phlebotomy tests. For clinical tests, all types of clinical tests were included to provide insights to decision making process and probable thoughts and concerns from GP practitioners during RTI consultations (Agniel et al., 2018). All clinical

investigation code list was sorted from the Test file and differentiated via Entity type into clinical check and clinical test. Clinical investigation episodes were identified from Clinical and Referral files.

7.2.5 Statistical approaches for predictor selection and model estimation

For this thesis, three analytical techniques were employed for prognostic modelling: random forests, penalized regression models, and Classification And Regression Trees (CART). CART was used to explore possible non-linear relationships and interactions between candidate predictors as well as result dissemination. The random forest was chosen primarily for predictor selection as shown by variable importance. Penalized regression models were fitted to provide complementary information to tree-based approaches during predictor selection. Conventional simple logistic regression models were also deployed for comparison purposes. An exploratory comparison of common machine learning approaches was conducted using the full dataset with 10-fold cross validation through the Classification Learner App in Matlab (The MathWorks, 2020) to provide complementary information for model specification.

7.2.5.1 Conventional regression modelling methods

The conventional approaches for binary outcome modelling are logistic regression, binomial regression with log link or Poisson regression with robust standard error (SE) (D Riley et al., 2019). Given that the outcome of interest being binary, logistic regression could be a candidate modelling method with estimated odds ratios being the main metric of interest. However, this approach is not straightforward for the number and type of candidate predictors for this project.

Firstly, if there are 30 candidate predictors, the space of all possible models lies between 30 to 1.07×10^9 theoretically. Manual selection based on predefined model specification criteria would be challenging. There are several available statistics for automated predictor selection such as the Akaike Information Criterion (AIC), similar to the equivalent adjusted R-squared in linear regression, but predictor

selection procedures like stepwise regressions (backward, forward and double directions) only offer limited benefit when the number of candidate predictors is more than 25 to 30 (Efron and Hastie, 2016a).

Secondly, if there is known multicollinearity among candidate variables, with several of them highly correlated, variable selection for logistic regression may not be straightforward.

Furthermore, logistic regression does not require a linear relationship between dependent and independent variables, but it does assume linearity between independent variables and the log odds which may not always be justifiable. It is very challenging to model multiple interactions using conventional regression modelling. Therefore, machine learning approaches that loosen specific statistical assumptions and enable the identification of strong, and non-linear associations from complex data are adopted for prediction modelling of this research (Chen and Asch, 2017).

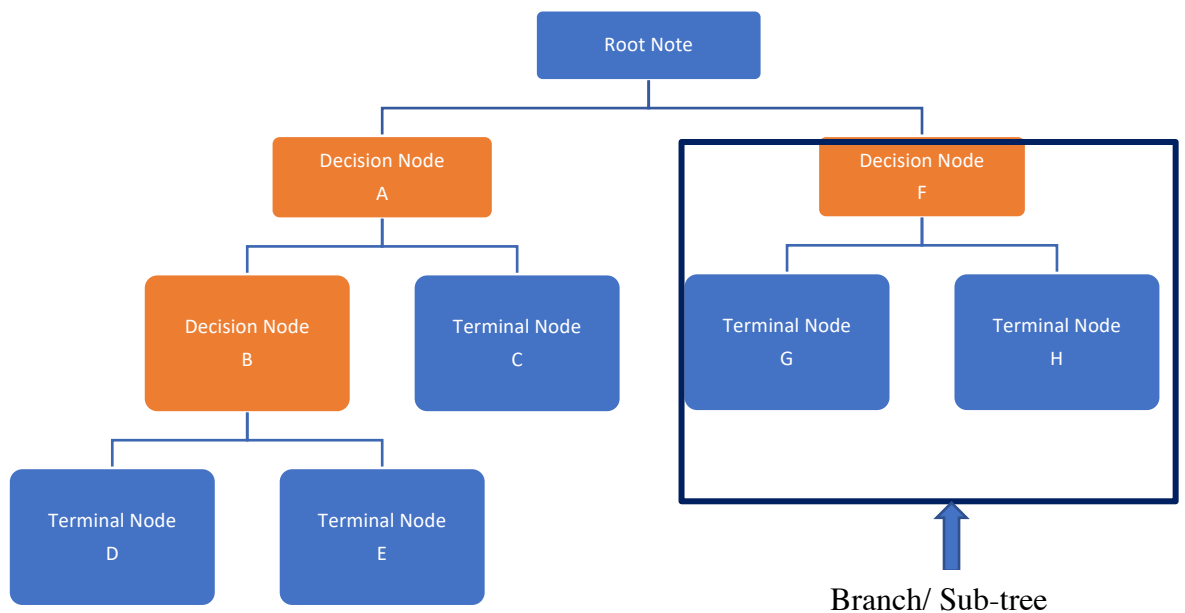
7.2.5.2 Classification and regression tree (CART) and Random forest

Tree-based modelling approaches like CART, random forest and gradient boosting are widely adopted with several benefits. Firstly, the output of the algorithms is comprehensible and allows for easy information visualization. The logic of tree construction is to employ a set of conditions to partition the total study sample into smaller but more homogenous groups in relation to the outcome of interest. The interpretation of a decision tree is intuitive as shown in Figure 7.3. The entire population sample (root node) was divided into two purer groups (A and F) according to splitting rules (discussed in detail below), then subsamples in these two nodes are split further down into sub-nodes referred to as decision nodes, also called parent nodes in relation to the sub-nodes (child-nodes) stemmed from it. The growth of the tree will terminate when stopping criteria are met. Then the bottom level of nodes that could not be split anymore are the terminal nodes- the leaves of the decision tree. When describing each leaf of the tree model, it may be considered as a multi-nested 'if-then' statement or combining all the conditions using 'and' that lie

on the pathways of that branch leading to that leaf (Figure 7.3: For terminal node D: all the conditions that resulted in decision nodes A and B together with the condition group D met after node B) (Kuhn and Johnson, 2013).

Moreover, this type of algorithms is equipped with the ability to map non-linear relationships and incorporate data of various kinds including continuous, discrete, skewed. Therefore, tree-based models may require less previous-knowledge and fewer clinical or statistical assumptions, and pre-processing of candidate predictors is not always essential, especially during data exploration. Since tree-based modelling algorithms are referred to as ‘greedy’ approaches which means that they only look for the best split for the next level and continue to move forward without looking a few steps ahead, this may result in over-fitting. Models based on a single tree may be less stable as a slight change of the data may lead to a different ‘tree’. Therefore, ensembled methods like random forest here was adopted to generate more robust results when compared with models based on single tree.

Additionally, these algorithms generally work better with categorical variables than continuous numerical variables as the tree grows through a recursive binary categorizing process. By assigning the sample population from a parent node into two independent, non-overlapping child nodes, prognostic information carried by continuous variables will be partially lost by such dichotomizations. To address potential disadvantages, pruning (discussed in detail below) as well as setting constraints on model parameters were used to deal with over fitting.



Note: A is the parent node of B and C, likewise B is the parent node of D and E.

CART

There are two major types of decision trees: if the dependent variable (outcome of interest) is categorical, then a classification tree is applied; if the target variable is continuous, then the model is called as regression tree. For this thesis, a classification decision tree is employed when referring CART.

How does the tree decide to split?

The decision tree grows ‘top-down’ from the node root to final leaves following the recursive binary splitting, therefore how the ‘tree’ decides to split is crucial determining prediction modelling accuracy. The ultimate goal for decision tree splitting is aiming to generate more pure next level of subgroups. Generally, statistics that measure the change of homogeneity or heterogeneity between parent node and its child nodes will provide suitable splitting criteria. For a classification tree, the Gini index, Chi-Square and Information gain are commonly adopted as metrics to describe and measure similarity or alternatively purity (Kuhn and Johnson, 2013).

The variables that the decision trees choose to split on are those that yield more homogenous sub-nodes comparing to the remaining predictors in terms of the splitting process taking place at that branch at the same level.

Here the *Titanic* data from ‘car’ package in R (Fox and Weisberg, 2019) was used to illustrate how these three statistics may be applied to decision tree when choosing variables to split on. There were 654 out of 2092 adult passengers who survived with 316 being women and 338 being men. Also, among these 654 survivors, 197 travelled on the first class and the rest 457 travelled on the other classes (second, third and crew). When the decision tree aims to predict the survival status of adult passengers, then it has to pick out the best predictor from candidate variables to split on based on the chosen statistic. In this case, candidate variables are sex (female or male) and travel class (first class or not), so the decision tree has to select one of them for the first level of decision node. Figure 7.4 gives a snapshot of two individual decision trees predicting survival status among adult passengers of Titanic.

As shown in

Table 7.2, sex is the better variable for the initial split comparing to the travel class as measured by Gini Index, Chi-Square and Information gain, which in line with previous studies to predict survival status of passengers on Titanic (Varian, 2014, Cicoria et al., 2014).

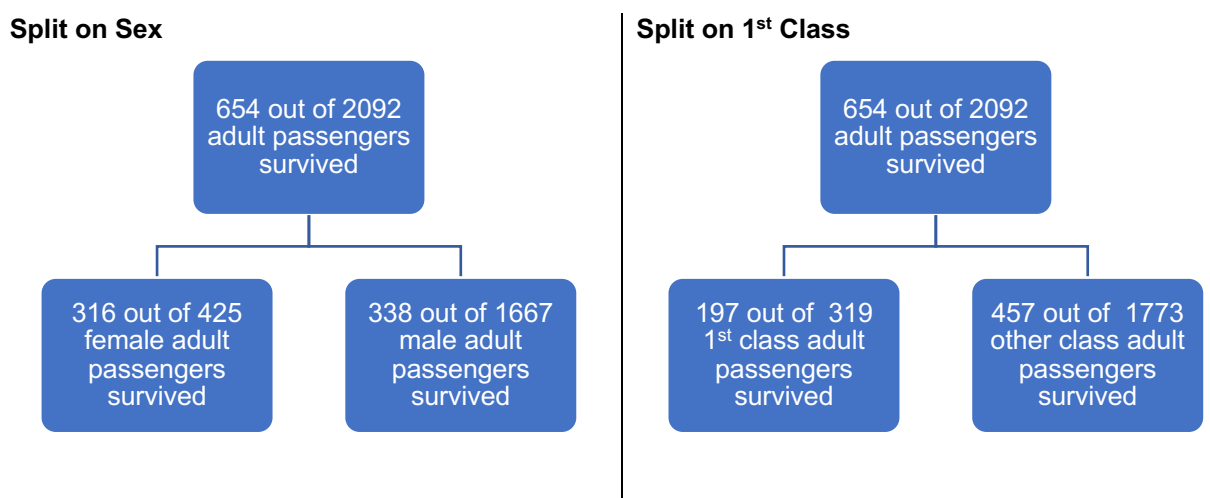


Figure 7.4: Illustration of decision tree splitting on sex vs travel class to predict survival status of adult passengers using *Titanic* data.

Table 7.2: Gini Index, Chi-Square and Information gain as splitting criteria for classification tree

Gini Index for a given node is $p_1(1-p_1) + p_2(1-p_2)$ since $p_1+p_2=1$ for binary class problem, then the calculation could be simplified as below. If the population is completely pure, the probability of that two random sample from the same class should be 1. Therefore, variable with the higher Gini value generally will be selected for tree splitting.

1. Calculate Gini for sub-nodes: sum of squared probabilities of 'success' and 'failure'
2. Calculate Gini for a given split using weighted Gini of each node of that split

Split on sex:

1. Gini for sub-node female: $(316/425)^2+(109/425)^2=0.62$
2. Gini for sub-node male: $(338/1667)^2+ (1329/1667)^2=0.68$
3. Weighted Gini: $(425/2092)*0.62+(1667/2092)*0.68=0.668$

Split on 1st Class:

1. Gini for sub-node 1st Class: $(197/319)^2+(122/319)^2=0.53$
2. Gini for sub-node other Class: $(457/1773)^2+(1316/1773)^2=0.62$
3. Weighted Gini: $(319/2029)*0.53+(1773/2029)*0.62=0.625$

Chi-Square measure the statistical difference between parent and child nodes, that is the bigger the difference is, the child nodes are more distinct from their parent nodes. The calculation for a give node is $((Actual - Expected)^2 / Expected)^{1/2}$, then the total sum of both classifications for that split.

Split on sex:

1. Actual number of female survivors vs non-survivors: 316 vs 109
2. The expected survival rate from parent node is $654/2029=0.32$
3. Expected number of female survivors vs non-survivors: $0.32*425=136$ vs $(1-0.32)*425=289$
4. Chi-Square for female survivors vs non-survivors: 15.43 vs 10.59
5. Calculate the chi-square for male survivors vs non-survivors: 8.46 vs 15.84
6. The total chi-square of splitting on sex is **50.32**

Split on 1st Class:

1. Actual number of survivors vs non-survivors from 1st Class: 197 vs 122
2. The expected survival rate from parent node is 0.32
3. Expected number of survivors vs non-survivors from 1st Class: 102 vs 217
4. Chi-Square for survivors vs non-survivors from 1st Class: 9.41 vs 6.45
5. Calculate the chi-square for survivors vs non-survivors from other class: 4.63 vs 3.18
6. The total chi-square of splitting on 1st Class is **23.67**

Information Gain which is derived from the information theory basically refers to the information needed to describe the characteristics of all objects assigned to one group. If the group is pure, then a simple sentence could be enough to cover the features that its participants share in common – that is little information is needed. The information gain could be simplified as 1-

Entropy where Entropy could be regarded as a measurement of the degree of impurity.

Therefore, the smaller the value of entropy is, the better the variable fit for the given split than the other candidate predictors.

The value of Entropy for a given node is: $-p\log_2 p - q\log_2 q$, therefore, the Entropy for a given split is calculated as

1. Calculate the entropy of the parent node.
2. Calculate the weighted entropy of the individual child nodes.

Split on sex:

1. Entropy for the parent node:
 $-(654/2092)*\log_2 (654/2092) -(1438/2092)*\log_2 (1438/2092)=0.89$
2. Entropy for female node:
 $-(316/425)*\log_2 (316/425) -(109/425)*\log_2 (109/425)=0.82$
3. Entropy for male node:
 $-(338/1667)*\log_2 (338/1667) -(1329/1667)*\log_2 (1329/1667)=0.73$
4. Entropy for the split on sex: weighted entropy of child nodes:
 $(425/2092)*0.82+(1667/2092)*0.73=\mathbf{0.76}$

Split on 1st Class:

1. Entropy for the parent node:
 $-(654/2092)*\log_2 (654/2092) -(1438/2092)*\log_2 (1438/2092)=0.89$
 2. Entropy for 1st Class:
 $-(197/319)*\log_2 (197/319) -(122/319)*\log_2 (122/319)=0.96$
 3. Entropy for other class:
 $-(457/1773)*\log_2 (457/1773) -(1316/1773)*\log_2 (1316/1773)=0.82$
 4. Entropy for the split on 1st Class: weighted entropy of child nodes:
 $(319/2092)*0.96+(1773/2092)*0.82=\mathbf{0.86}$
-

How to avoid over-fitting during CART modelling process?

There are two major stages at which constraint procedures could be applied during decision tree modelling: either setting constraint parameters on the tree itself or pruning the tree after it has grown.

Setting constraints on the tree, defines the size of the tree by adding requirements to the sample size of a node, the minimum sample size for a split; a leaf (terminal node), the minimum sample size to terminate splitting process; height of the tree, the maximum depths or the maximum levels to split in return determine the maximum number of leaves (for classification trees, the maximum number of leaves should be twice of the depths of the trees); the maximum number of predictors for split.

Various considerations should be given when selecting among these constraint parameters depending on the specific case. For example, if there are class imbalance problems, then a small sample size of terminal nodes would be chosen, so that the overall model will not be dominated by the majority class and underlying relationships between predictors will not be masked by such imbalances.

In contrast to setting constraint parameters and fixing the tree size within a range, pruning starts from the bottom level and moves backward after a decision tree grows to a large depth. By removing the leaves which give suboptimal prediction performance in comparison with the previous splitting, pruning is a complementary method to the natural 'greedy' approach of decision tree growing since it will allow insights to a few steps ahead and tailor the model to sensible sizes for specific research scenarios.

Random forest

Random forest could be viewed as ensembled decision tree adopting the bagging concept. As shown in Figure 7.5, the bagging technique starts by random sampling with replacement (bootstrap sampling) from the original data to create multiple sub-data sets. The new sub-data sets could have a proportion of the predictors as well as the total sample size. Then a prediction model is generated for each specific sub-data

set (M_1 to M_n), and final model (M^*) is ensembled based on the mean, median or other modelling statistics of these models depends on the specific research question to answer. Meanwhile, the data is not sampled for each model is referred to as the ‘out of bag’ (OOB) sample, which may be used as test data to estimate the errors of these models. Errors estimated on the OOB samples are called as ‘out of bag’ errors. Conventionally, the algorithms sampled two thirds of the total data.

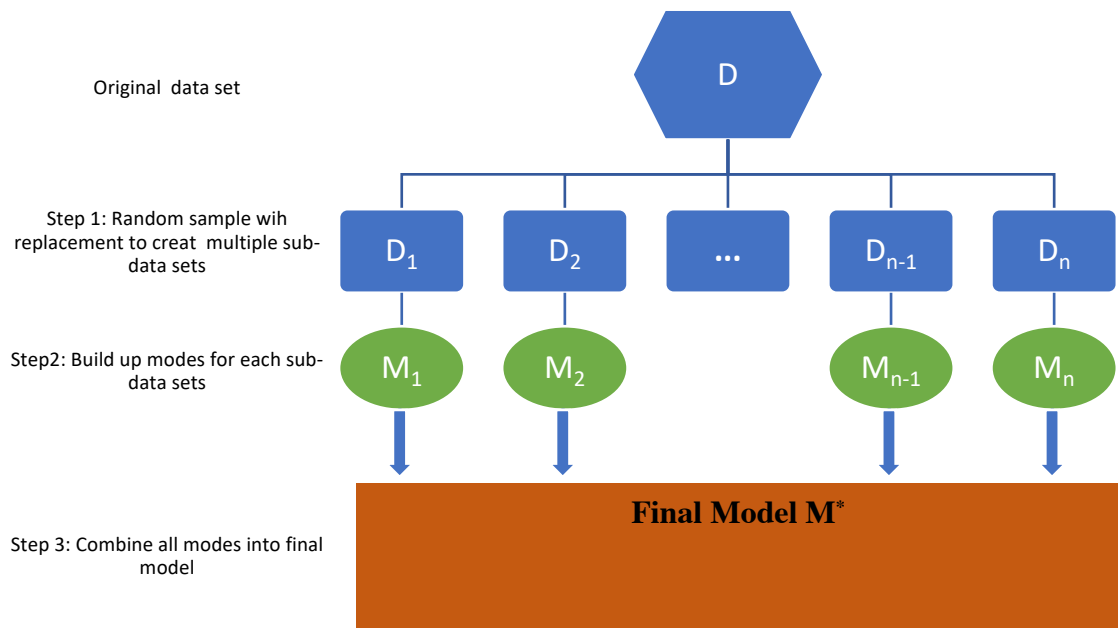


Figure 7.5: Illustration of bagging process

When it comes to decision trees, multiple trees are planted using the bagging technique. That is if the total sample size is M , then m cases will be randomly selected with replacement out of it ($m < M$) to create sub-sets of data. Likewise, if the total number of candidate predictors is N , n variables are chosen randomly for the best split to grow each tree. The value of n remains to be constant during each random sampling and the square root of the number of candidate predictors (\sqrt{N}) is generally recommended for n (Breiman, 2001). After each tree is grown to its extreme depth without constraints or pruning, a forest is formed as so called ‘random forest’.

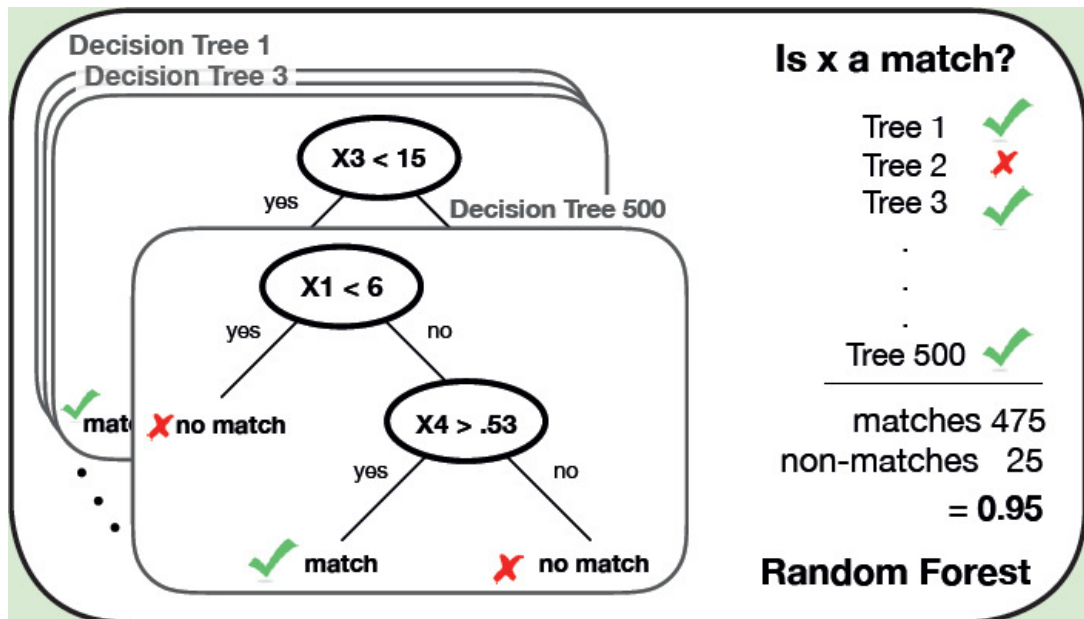


Figure 7.6: Illustration of random forest for classification trees (Carriquiry et al., 2019).

Figure 7.6 demonstrates the voting process for a specific predictor X. In a forest, hundreds of trees (500 here) are fitted into bootstrapped samples of the training data, in which the tree considers a random subset of the features at each candidate split. The output from a random forest is the probability of class membership for each item by aggregating the results from each tree. Therefore, predictor X is considered to be a match given its probability of class membership being 95% as illustrated above.

The result of random forest could be the major votes for classification trees which is displayed as variable importance. Such variable importance will provide information on variable selection which could be used for dimension reduction. Since random forest aggregates the results from multiple decision trees, its performance is more stable than a single tree and tends to deal with non-informative predictors in terms of the outcome of interest better than other machine learning methods e.g. neural networks (Kuhn and Johnson, 2013).

Subject to the generic limitations of tree-based approaches, random forest works better with classification problems than regression ones. Also, it does not present detailed process in terms of how the algorithms work more of like a ‘black box’, so that users have little control during modelling. As a result, more than one random

seed and different parameters are set in this study to ensure the reproducibility of the results meanwhile to avoid getting various results by random chance especially for key variables. For the initial number of trees that a random forest should grow, there is no generic rule but varies case by case. In this prognostic study, the random forest begins with 50 trees, and adjustment is made using the error rates (OOB error rate and misclassification error rate) vs number of trees plot as reference.

7.2.5.3 Penalized regression: Lasso, Ridge and Elastic net

As noted above, simple linear and logistic regression approaches generally do not fit for big data for prognostic studies as but, they remain classic tools that foster an array of modified regression approaches to explore the relationship between single/multiple dependent variable(s) and independent variable(s). Figure 7.7 lists out the common regression models based on the key metrics they emphasise.

Basically, simple regression algorithms aim to fit a line or curve which has the minimum distance or deviance from observed data points in the model development dataset. Such minimum distances are measured by the least-squares of vertical distance between each data points and fitted line (predicted values) in simple linear regression, so that the effect of positive and negative values is eliminated from the measurement. Likewise, in simple logistic regression, the model is estimated by maximizing the likelihood of observing the individual data points in the dataset using logit function (sigmoid function) so that the binary outcome (yes or no) could be transferred to a continuum of probabilities ranging from 0 to 1 (Mitchell, 1997). Once the model is fixed, then the average effects of individual variables as expressed by the magnitude of coefficients are quantified. Apart from the predetermined qualities of dependent and explanatory variables, the choice between regression modelling approaches mainly lie in which regression technique will generate a good model. Here, linear regression is adopted to illustrate the advantages of penalized regression in comparison with the rest of this family for prognostic research using big data. Because there is a separate error item as denoted by ϵ in the simple linear regression equation $y = \beta_0 + \beta_1 x + \epsilon$, which is more intuitive to understand than logistic regression where the error item is integrated into the model estimation.

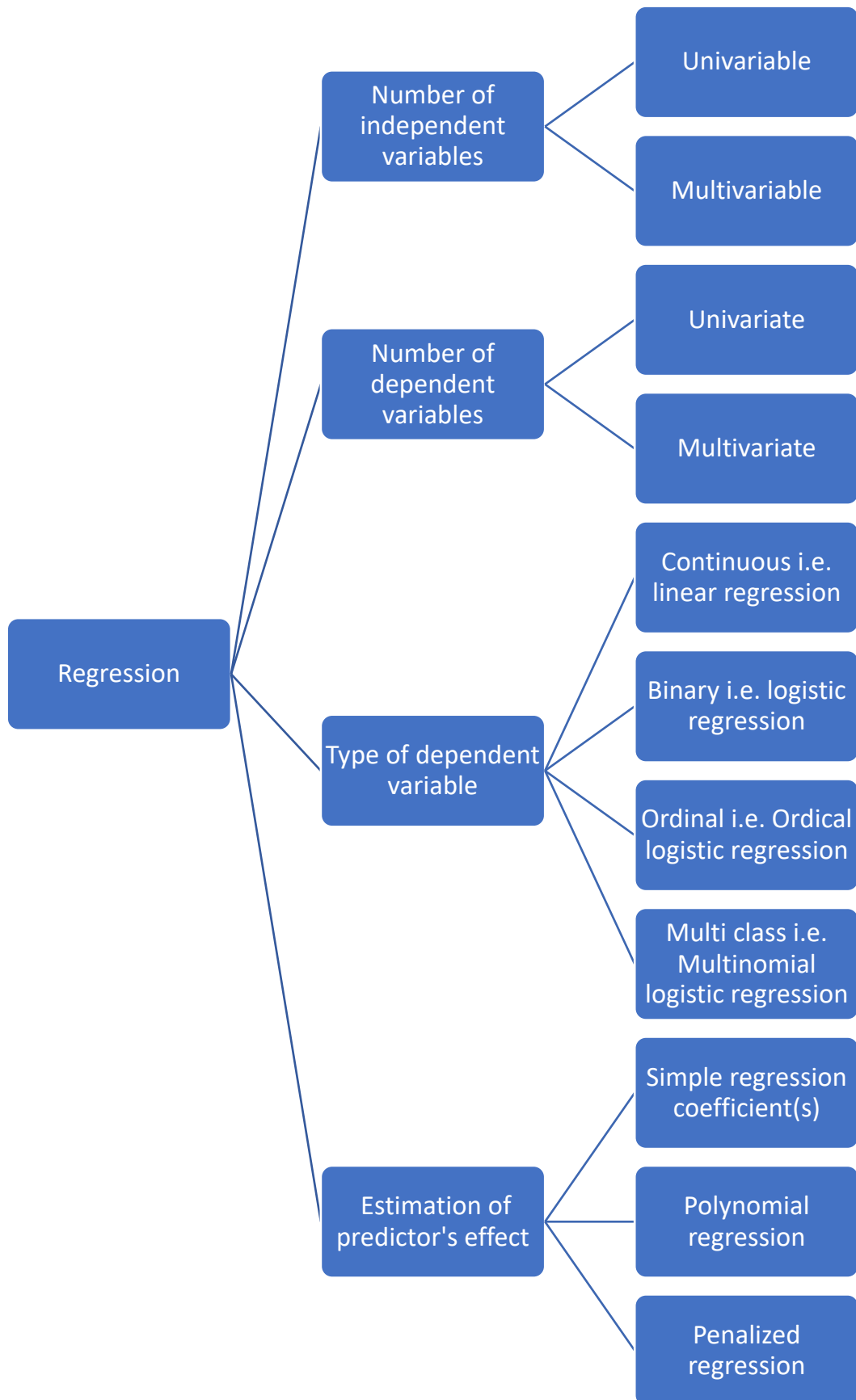


Figure 7.7: Common regression models based on key metrics

The line of best fit

As mentioned before, for simple linear regression, the best line fitted to the data is the one closest to all the observation points. That is the model endeavour to minimize the differences between predicted and actual values, such differences are termed as errors. So, the main questions for linear regression model development could be partitioned into two major aspects: where these errors derive and how to minimize them.

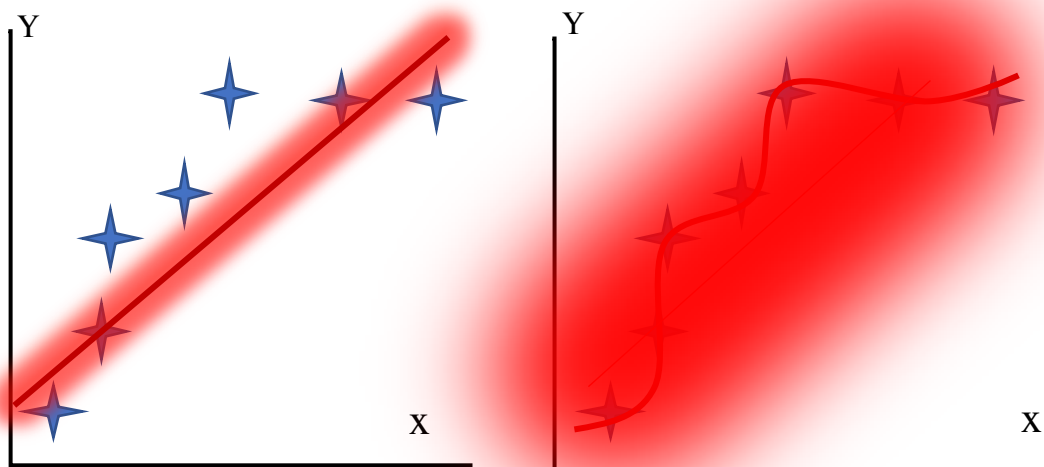
Before minimizing the errors of the model, sourcing and quantifying the component of the errors should be done. The total error of a linear regression is calculated as sum of the distances also called as residuals between the observed values and predicted values on the fitted line. In order to prevent the offsets between positive deviance and negative deviance, both absolute value and squared value could be considered as measurements of the total deviance. In conventional practice, sum of squared residuals is used to measure the total error of a linear regression model since it penalizes higher error values more which differentiate the contributions from big errors and small errors. Therefore, the sum of squared errors (SSE) is calculated as $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ where y_i is the observed value and \hat{y}_i is the predicted value, n refers to the sample size of the model development sample size (or the number of rows in the training data set). Also, the average error of the given data set as denoted by mean squared error (MSE) could be generated as SSE/n , that is

$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$. If the estimation of MSE is aimed, the most simplified scenario would be the data points are independent from each other statistically and the theoretical mean of the residuals is zero with a constant variance of σ^2 , then the expected MSE could be decomposed as:

$E[MSE] = \sigma^2 + (\text{Model bias})^2 + \text{Model variance}$ (Kuhn and Johnson, 2013). So that the source of overall model error is distilled into model bias and model variance.

Model bias reflects how much developed model deviate from the actual data whereas model variance refers to the stability of the model. In another word, if the model adopted as many informative variables as possible to describe the data set, then the bias would be low as the fitted line almost sketches the joint line of the observation

points. However, this line is very easily affected by additional unseen data points even if those points stay closely to this line not to mention outliers, that is, the possible models lie in anywhere and take on any shape within the shaded area as shown on the right panel of Figure 7.8. On contrary, if the model is built to be simple as portrait on the left panel of Figure 7.8, the total error of which is greater, thus even if an outlier will not drive the line towards its direction easily given its marginal error contribution to the overall deviance. Therefore, the simple model has less variance, and the space of possible models is niched as shown by the shadowed area around the straight line, but it fails to reflect the characteristics of the population as accurate as the high variance model. This will introduce a key question in regression modelling: what is a good model?



High bias low variance (Underfit)

model

$$\beta_0 + \beta_1 X$$

Low bias high variance (Overfit)

model

$$\beta_0 + \beta_1 X + \beta_2 X + \dots + \beta_j X$$

Figure 7.8: Illustration of under fit vs over fit model

What is a good model?

Accurately predicting the outcome of interest without overfitting leads to a stable model that may have better potential applications to wider population, further gauge

the usefulness of the prediction model. Given the context of regression modelling, the optimal model would be minimizing the model error while avoiding unnecessary model complexity. As illustrated by Figure 7.9, the trade-off relationship between the two essential components (model bias and variance) of total model error, the optimal model is located around the minimum total error which is one of the main targets of regression modelling.

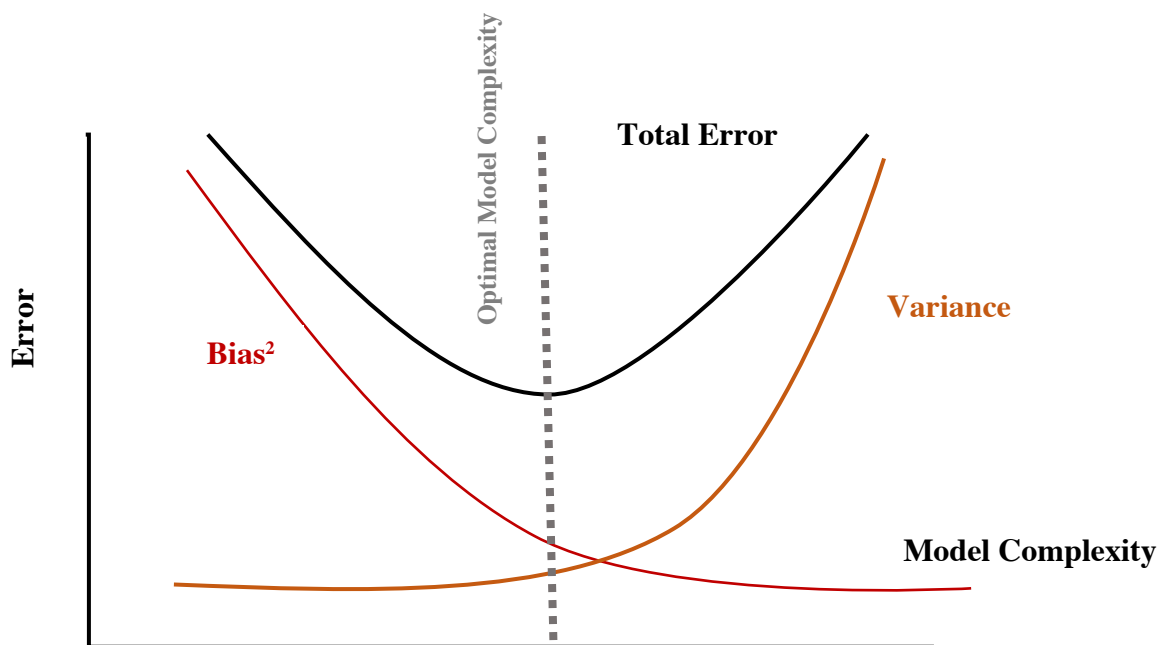
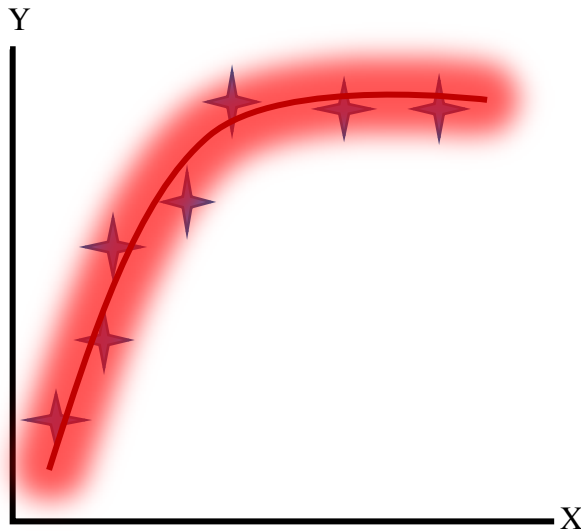


Figure 7.9: Trade-off between model variance and bias (Huilgol, 2020)

To address underfitting, simply adding more relevant variables to the model would be helpful to reduce its bias. Whereas to overcome overfitting, either the model reduces its complexity by using fewer predictors which requires pre-knowledge or minimizes the magnitude of the coefficients which indicate the contributions of individual predictors to the prediction model. During initial modelling stage, determining the prognostic values of all candidate variables is quite unlikely. Therefore, reducing the magnitudes of coefficients meanwhile to quantifying the emphasis given to each predictor to inform predictor selection is a sensible approach. In return, the optimal model could alternatively be interpreted as a less curvy line (in comparison with the overfitting model) at an acceptable bias level with the main objective being minimizing the model error as illustrated by Figure 7.10. By

smoothing the overfitting line, the bias of the model is inevitably increased, but the magnitude of the coefficients of the model is shrunken as shown by the shadowed area along the line. Now, the crucial question becomes how to realize this objective.



'Just right' model

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

Figure 7.10: Illustration of the optimal regression model

Regularization refers to adding penalty terms to the sum of squared errors of the regression function, so that the sum of errors the algorithms aim to minimize is changed from sum of squared error $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

into

$SSE + \text{penalty term}$,

where the penalty terms are usually the transformations of the magnitude of the model coefficients. Then the total error that the model targeted to minimize in order to get the best fit line has become bigger, plus the penalty term is generated from the coefficient magnitude, thus the final model regularizes the magnitude of coefficients meanwhile minimizes the total error of the model. The key is how to assign the penalty term to the model. Here Ridge regression and least absolute shrinkage and selection operator (lasso) regression are initially adopted as common regularization techniques to explain how penalized regression works followed by their hybrid function elastic net regression.

Ridge and Lasso regressions

In ridge regression, the minimization target is set as SSE plus the square of coefficient magnitude (Kuhn and Johnson, 2013) which is a second-order penalty as indicated by ‘L₂’ below:

$$SSE_{L_2} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Whereas, in lasso regression, the penalty term is the absolute value of the magnitude of coefficients (Kuhn and Johnson, 2013) which is L₁ penalty as denoted below:

$$SSE_{L_1} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Even if the difference between ridge and lasso regression functions seems trivial, this exactly makes an important distinction in practical utilization. Before proceeding with how the penalty terms affect model behaviour, the parameter λ which balances the emphasis between sum of squared error and sum of squared coefficient magnitude when minimizing SSE_{L₁} or SSE_{L₂} as a whole should be explained.

Technically λ could take on any non-negative value: if λ equals 0, then the model is the same as simple regression model; if it takes a positive value of infinite ∞ , then the coefficients in the model will be 0 (lasso) or infinitely close to 0 (ridge).

Therefore, the final λ for a given model should lie in somewhere between 0 and infinite. The widely accepted method to find the ideal λ is repeatedly testing a series values within a predefined range using cross validation (10-fold for this thesis). And the one yields the best performance as measured by chosen statistics i.e. MSE or misclassification error will be selected (Efron and Hastie, 2016b) as shown by the left vertical dot line on the right panel of the graph (Figure 7.11). Sometimes, λ plus one SE is recommended as a more conservative practice.

After working out how to allocate the value of λ , there are differences between ridge and lasso regressions during the regularization implementation. As illustrated below Figure 7.12, minimization targets should start around the places where the contours touch the penalty region. Apparently, lasso is able to pull the coefficients to absolute zero which is absent from ridge. This will make lasso suitable for automatic variable selection by shrinking certain variable coefficients into zero, leaving a narrow option spectrum. Although the ridge method is not suitable for variable selection purposes, it retains all the variables in the model with various predictive values. This is

particularly powerful to deal with multicollinearity among predictors by quantifying the individual contribution of each variable to distil the predictive value of several highly correlated predictors without losing information in comparison with lasso (Kuhn and Johnson, 2013).

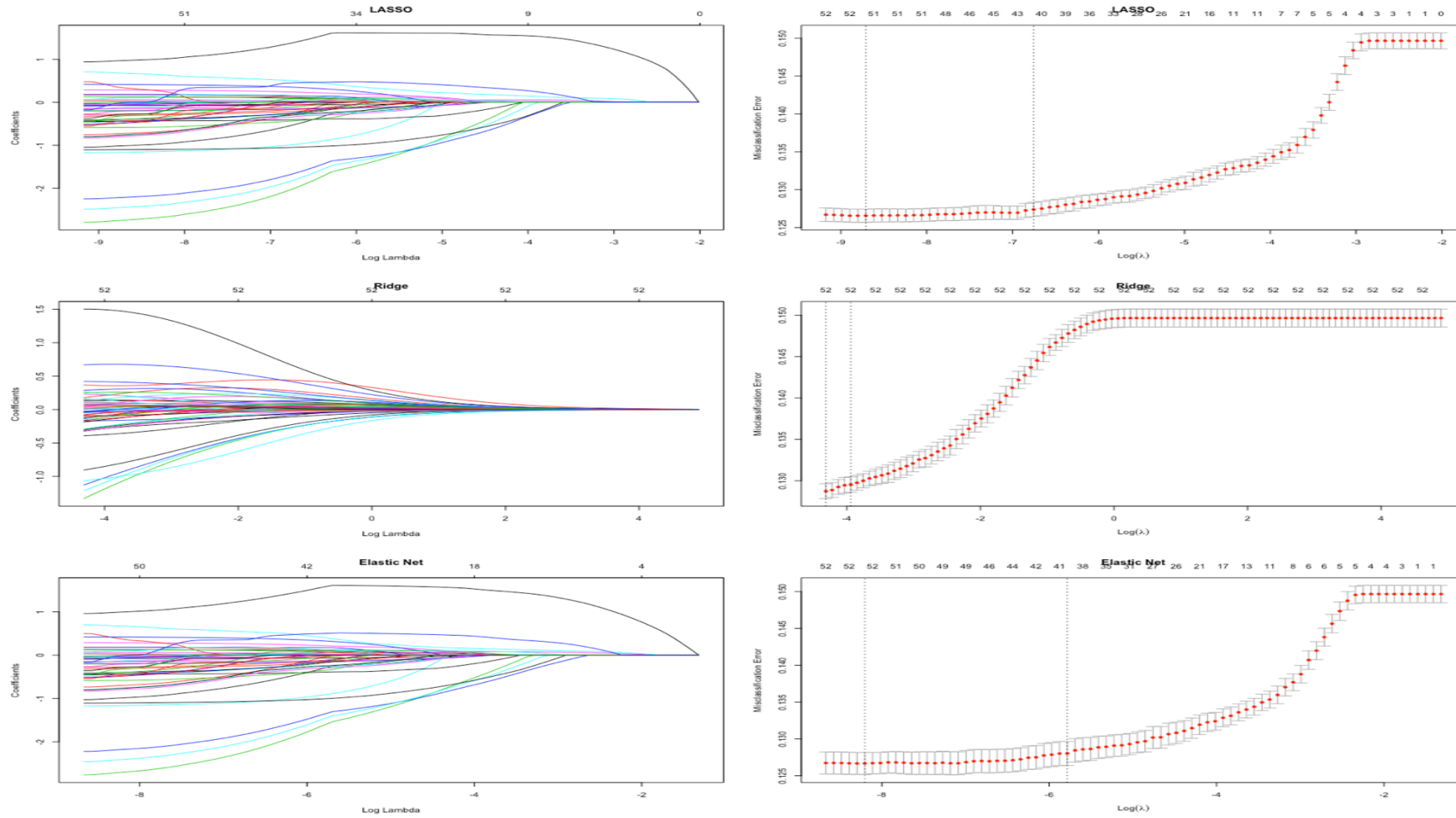
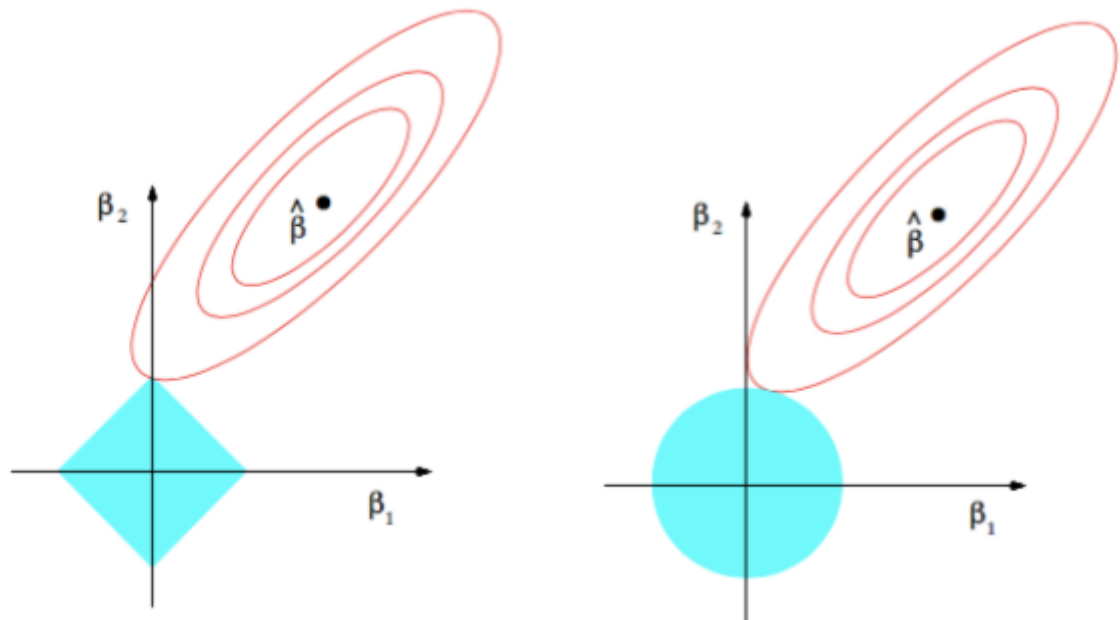


Figure 7.11: Lambda for three penalize regressions (full model)

Lasso regression

Ridge regression



Red contour: least squared error term SSE_{L_2} (left: lasso) SSE_{L_1} (right: ridge)

Blue shape: the type of penalty (Left: L_2 vs Right: L_1)

Black point: $\hat{\beta}$ refers to the un-restricted least square estimation of β .

Figure 7.12: Exemplar of difference between ridge and lasso regression (Efron and Hastie, 2016b)

Elastic net regression

Elastic net regression is considered as the hybrid function of Ridge and Lasso regressions with the penalty term being: $(\lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2)$ where $\lambda_1 + \lambda_2 = 1$. That is, when the algorithm minimizes the penalty term as a whole, λ_1 and λ_2 quantify the weights assigned to these two shrinkage techniques. If $\lambda_1 = 1$ and $\lambda_2 = 0$, it performs Lasso regression; if $\lambda_1 = 0$ and $\lambda_2 = 1$, it turns into Ridge regression. Given that Lasso regression has the capability to shrink certain variable estimations into zero, such exercise maybe not as stable as Ridge regression since it can pick out one variable

randomly especially when a group of predictors are highly correlated. Therefore, Elastic net regression adopts the practical advantages of Lasso and Ridge's stability.

7.2.6 Final model selection

Final model selection will incorporate both statistical and clinical significances into consideration since the main purpose for this prognostic modelling study is to assist clinical decision at point of care. Therefore, caution is required during the interpretation of statistical output because EHRs were not structured for research purposes, rather for daily clinical care; consequently, certain predictors may not be merely the physiological surrogates of health conditions but might reflect the outcome of interactions between patients and health care system.

During variable selection using the random forest model, Gini-based variable importance estimates were reported. The number of trees were selected based on the OOB misclassification errors after random forest models fitted with varying number of trees (*n_{tree}*: 50, 100 and 150). Variable importance generated from different numbers of variables randomly sampled when creating the trees were reported. The number of variables fitted into random forest models began with five in increments of five to 30 in order to control the complexity of learning function was adopted (Hothorn et al., 2006, Strobl et al., 2008). Given that Gini-based variable importance estimates are prone to bias towards numerical and categorical variables with more categories (Strobl et al., 2007), variable selection process and further variable binning will not be algorithms driven only, rather based on comparisons between modelling approaches as well as clinical guidelines at ease of clinical use.

7.2.7 Statistical software and main packages

Analysis were performed in R program version 3.6.3 (R Core Team, 2020). The 'caret' package was used for analysis (Kuhn, 2020), this was supplemented by the 'rpart' package (Therneau and Atkinson, 2019) for CART modelling, 'randomForest' (Liaw and Wiener, 2002) and 'glmnet' (Friedman et al., 2010) packages for random forest and penalized regression respectively. The

'vip'(Greenwell et al., 2020) and 'ggplot2' (Wickham, 2016b) packages were used to construct variable importance plots. 'ROCR' (Sing et al., 2005) and 'pROC' (Robin et al., 2011) packages were used to assess model performance.

7.3 Data governance approval

The research protocol for this study was submitted to and approved by the Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC), Protocol 16_020. All patient medical records were anonymised before data received by researchers.

Chapter Eight : Clinical prediction model development for adult RTI patients who reconsulted with pneumonia within 30 days (Results, Discussion and Conclusion)

This chapter presents the results, discussion and conclusion sections of the prediction modelling study of the thesis following on from the methods outlined in the previous chapter.

8.1 Results

In this section, results of the prediction model development are presented in four sections: the descriptive statistics for study sample, the variable selection for the final model, the final model development and performance evaluation, followed by subgroup analysis.

8.1.1 Descriptive statistics of study population

The analysis included 108,842 patients who presented with RTI consultations in primary care between 1st January 2002 and 31st December 2017, 16,289 of whom reconsulted with pneumonia within 30 days after the RTI index date. All descriptive statistics presented in this section aimed to present a basic description of the data with estimations reserved for model development study.

Table 8.1 presents the descriptive characteristics of the sample. Overall, patients with pneumonia were older than RTI patients who did not develop pneumonia with mean age being 65.5 years and 49.2 years respectively. Women accounted for a high proportion in both groups. The seasonal distribution was similar between the two groups. The majority of study population were non-smokers (66.1%) with similar proportions of current smokers found in both groups. Slightly higher proportions in underweight and healthy weight categories were found in patients with pneumonia group.

Generally, the health status of pneumonia patients was more frail comparing to that of non-pneumonia cohort since larger proportions of mild to severe frail patients were identified from pneumonia cohort, meanwhile about 60% of non-pneumonia patients were presented being fit as measured by electronic frailty category as shown in Table 8.2.

Pneumonia patients showed higher proportions of the chronic conditions listed in Table 8.3 compared to non-pneumonia patients. There were 53.7% of non-pneumonia patients who were free of pre-diagnosed chronic conditions (chronic respiratory disease, chronic heart disease, chronic kidney disease, chronic liver disease, chronic neurological disease, diabetes, immune system condition), but 65.3% of pneumonia patients presented with comorbidities.

Disparity of general health status as measured by frailty and comorbidity was noticed as shown in Table A 10, 76.8% of patients without any included comorbidities were classified as non-frail and 18.5% were evaluated being mildly frail. Meanwhile, approximately 32% of patients with multi-comorbidities were considered to have mild or moderate frailty.

For initial RTI consultations, there were more pneumonia cases presented with LRTIs labelled as chest infection (48.8%) whereas non-pneumonia patients were presented with more proportions of URTI diagnoses as shown in Table 8.4.

Apart from 'clinical check' that both groups received more than 99% of medical management, high proportions of cases were found to have experienced medical interventions for the rest remaining categories of medial management including antibiotic treatment or clinical investigative tests in Table 8.5.

Disparity was noticed for antibiotic prescription both on the index date as well as before pneumonia re-consultation date for any other indications. There were significant higher proportions of pneumonia patients who received flu (60.9%) and pneumococcal (51.5%) vaccinations.

Table 8.1: Descriptive statistics of demographic characteristics for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.

	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)
Age		
Mean (SD)	49.2 (19.7)	65.5 (19.3)
Age group		
16-35	26,518 (28.7)	1,446 (8.9)
36-45	15,640 (16.9)	1,598 (9.8)
46-55	14,616 (15.8)	1,725 (10.6)
56-65	13,880 (15.0)	2,480 (15.2)
66-75	11,607 (12.5)	3,020 (18.5)
76-85	7,669 (8.3)	3,549 (21.8)
86 and above	2,623 (2.8)	2,471 (15.2)
Sex		
Female	56,936 (61.5)	8,755 (53.7)
Male	35,617 (38.5)	7,534 (46.3)
Season		
Autumn	21,365 (23.1)	3,865 (23.7)
Spring	23,361 (25.2)	3,971 (24.4)
Summer	16,965 (18.3)	2,846 (17.5)
Winter	30,862 (33.3)	5,607 (34.4)
Smoking Status		
Non-smoke	62,286 (67.3)	9,612 (59.0)
Ex-smoker	12,810 (13.8)	3,564 (21.9)
Current smoker	17,457 (18.9)	3,113 (19.1)
BMI (Kg/m²)		
Under weight (<18.5)	2,120 (2.3)	749 (4.6)
Healthy weight (18.5-24.9)	29,454 (31.8)	5,588 (34.3)
Overweight (25.0-29.9)	26,282 (28.4)	4,546 (27.9)
Obese (30.0-34.9)	12,853 (13.9)	2,062 (12.7)
Severe obesity (35.0-35.9)	4,899 (5.3)	743 (4.6)
Morbid obesity (40+)	2,739 (3.0)	458 (2.8)
Not recorded	14,206 (15.3)	2,143 (13.2)

Table 8.2: Descriptive statistics for frailty for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.

	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)
Frailty Category		
Fit	54,777 (59.2)	5,468 (33.6)
Mild	23,527 (25.4)	4,886 (30.0)
Moderate	9,734 (10.5)	3,540 (21.7)
Severe	4,515 (4.9)	2,395 (14.7)
eFrailty index		
Mean (SD)	0.13 (0.11)	0.20 (0.13)

Table 8.3: Descriptive statistics of chronic conditions including co-morbidity for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.

	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)
Chronic Respiratory Disease	28,265 (30.5)	6,126 (37.6)
Chronic Heart Disease	7,682 (8.3)	3,301 (20.3)
Chronic Kidney Disease	5,050 (5.5)	2,500 (15.3)
Chronic Liver Disease	401 (0.4)	177 (1.1)
Chronic Neurological Disease	4,358 (4.7)	2,811 (1.1)
Diabetes	6,372 (6.9)	2,284 (14.0)
Immune System Condition	13,621 (14.7)	2,962 (18.2)
Comorbidity		
No Comorbidity	49,662 (53.7)	5,651 (34.7)
One Comorbidity	26,645 (28.8)	4,984 (30.6)
Multi-Comorbidity	16,246 (17.6)	5,654 (34.7)
Cancer	4,851 (5.2)	2,284 (14.0)
Peptic Ulcer	2,532 (2.7)	1,044 (6.4)
Peripheral Vascular Disease	1,742 (1.9)	1,062 (6.5)
Hemiplegia	172 (0.2)	128 (0.8)
Charlson Comorbidity Count		
0	52,182 (56.4)	5,142 (31.6)
1	28,523 (30.8)	5,027 (30.9)
2	7,434 (8.0)	2,964 (18.2)
3	2,866 (3.1)	1,734 (10.6)
4	1,046 (1.1)	878 (5.4)
5	364 (0.4)	371 (2.3)
6	138 (0.1)	173 (1.1)

Table 8.4: Descriptive statistics of initial RTIs for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.

	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)
Cold/Influenza/URTI	16,555 (17.9)	1,633 (10.1)
Sore throat	20,230 (21.9)	466 (2.9)
Rhinosinusitis	7,088 (7.7)	138 (0.8)
Otitis media	5,202 (5.6)	78 (0.5)
Cough	37,211 (40.2)	6,483 (39.8)
Chest Infection	9,461 (10.2)	7,942 (48.8)

Table 8.5: Descriptive statistics of medical management for non-pneumonia patients and pneumonia patients. Figures are frequencies (column percentages) except where indicated.

	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)
Antibiotic prescription on the RTI index date	49,591 (53.6)	9,114 (56.0)
Antibiotic prescription in the following 30 days after initial RTI consultations	54,008 (58.4)	9,897 (60.8)
Asthma drug	10,108 (10.9)	935 (5.7)
Flu vaccination	28,347 (30.6)	9,009 (55.3)
Pneumococcal vaccination	23,367 (25.2)	8,392 (51.5)
Clinical test	15,144 (16.4)	3,244 (19.9)
Clinical check	92,444 (99.9)	16,232 (99.7)
Hospital admission in previous year	410 (0.4)	236 (1.4)

8.1.2 Variable selection

An iterative approach was adopted during variable selection. That is, all candidate variables were fitted into random forest models (Figure 8.3), simple logistic regression (with backward elimination) and penalized regression models (Lasso, Ridge and Elastic net). Elimination and refinement of predictors for further analysis were performed based on estimates from these algorithms together with consideration of clinical relevance.

Initially, random forest models were fitted with varying number of trees (*ntree*: 50, 100 and 150) shown as Figure A 1 (A, refers to Appendix). Given that the OOB misclassification errors were similar, subsequent models were fitted using ‘*ntree*’ value of 50. Next, *mtry* (number of variables randomly sampled when creating the tree) with 5 to 30 in increments of 5 to control the complexity of learning function was adopted and generated six variable importance plots with the top 20 important variables illustrated in Figure 8.1.

Second, all 36 variables were fitted into simple logistic regression. After backward stepwise selection, 32 variables were retained in the final model with antibiotic prescription in the following 30 days after initial RTI consultations, eFrailty index, Charlson comorbidities of ‘cancer’ and ‘hemiplegia’ being removed. Variable importance of simple logistic regression was ranked based on the absolute value of the t (or z) statistics which are the parameter estimates divided by their SEs. Variable importance plot of simple logistic regression together with those from penalized regressions are presented in Figure 8.3.

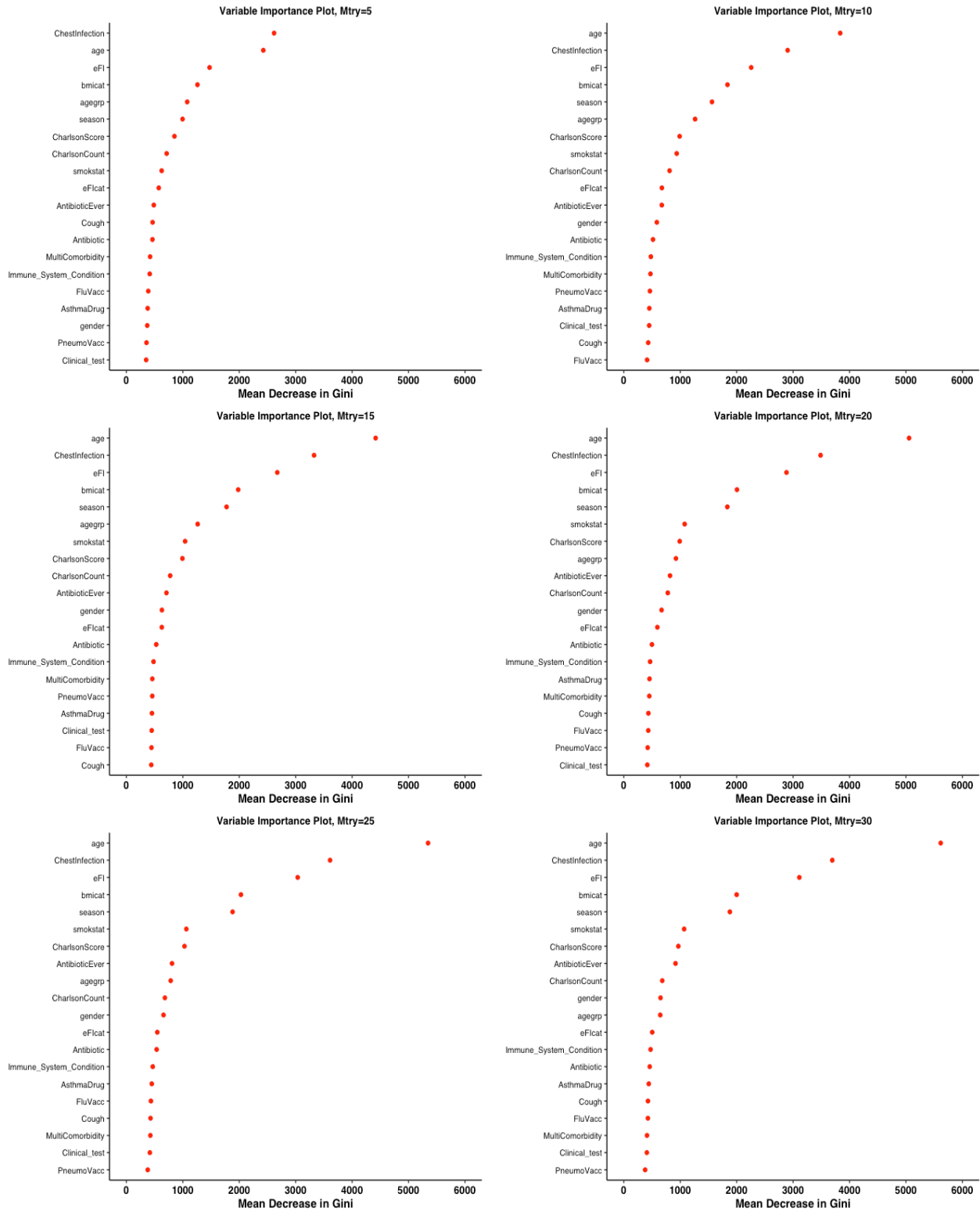
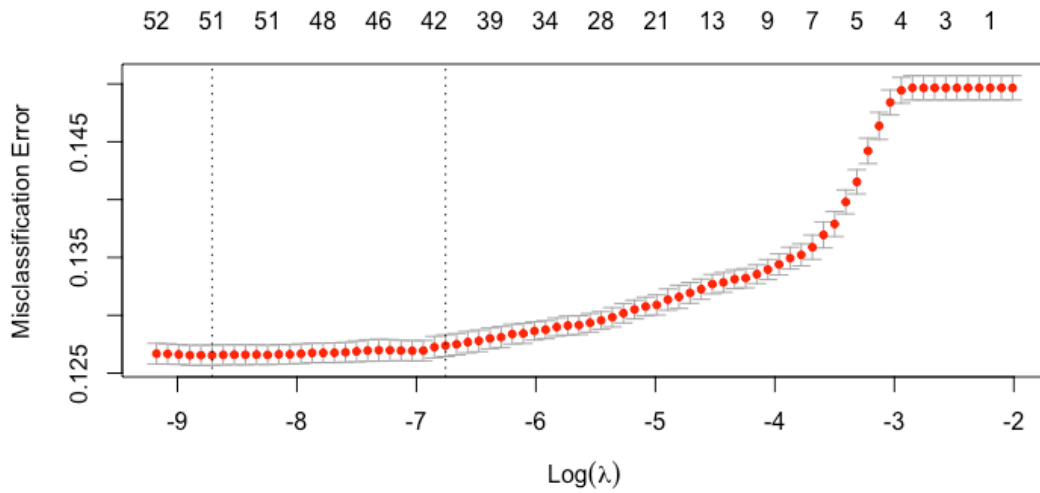


Figure 8.1: Variable importance results by random forest models with Mtry 5-30 (5 increments) for full model

Finally, penalized regression models were fitted using the lasso, ridge and elastic net procedures as outlined in the methods section of the previous chapter. The lasso model was fitted using the binomial family and an alpha value being one. The best lambda was selected using 10-fold cross validation and the cross-validation curve shows the upper and lower standard deviation along the sequence.



Minimum lambda: 0.0002; minimum lambda plus its one SE: 0.001

Figure 8.2: Cross validation curve of lambda for lasso regression

The two lines represent the lambda of minimum classification error and the minimal lambda plus its one SE (which is more regularized than minimum lambda).

Regression coefficients could then be evaluated at minimum lambda or minimum lambda plus one SE. To analyse the best model further, the ‘glmnet’ model was refitted using the minimum lambda in order to estimate the regularized regression model. Variable importance was ranked according to the absolute value of coefficients of penalized regression model. Both absolute values and directions of the coefficients are reported in Table A 11. Ridge regression models were evaluated in a comparable way Figure A 2.

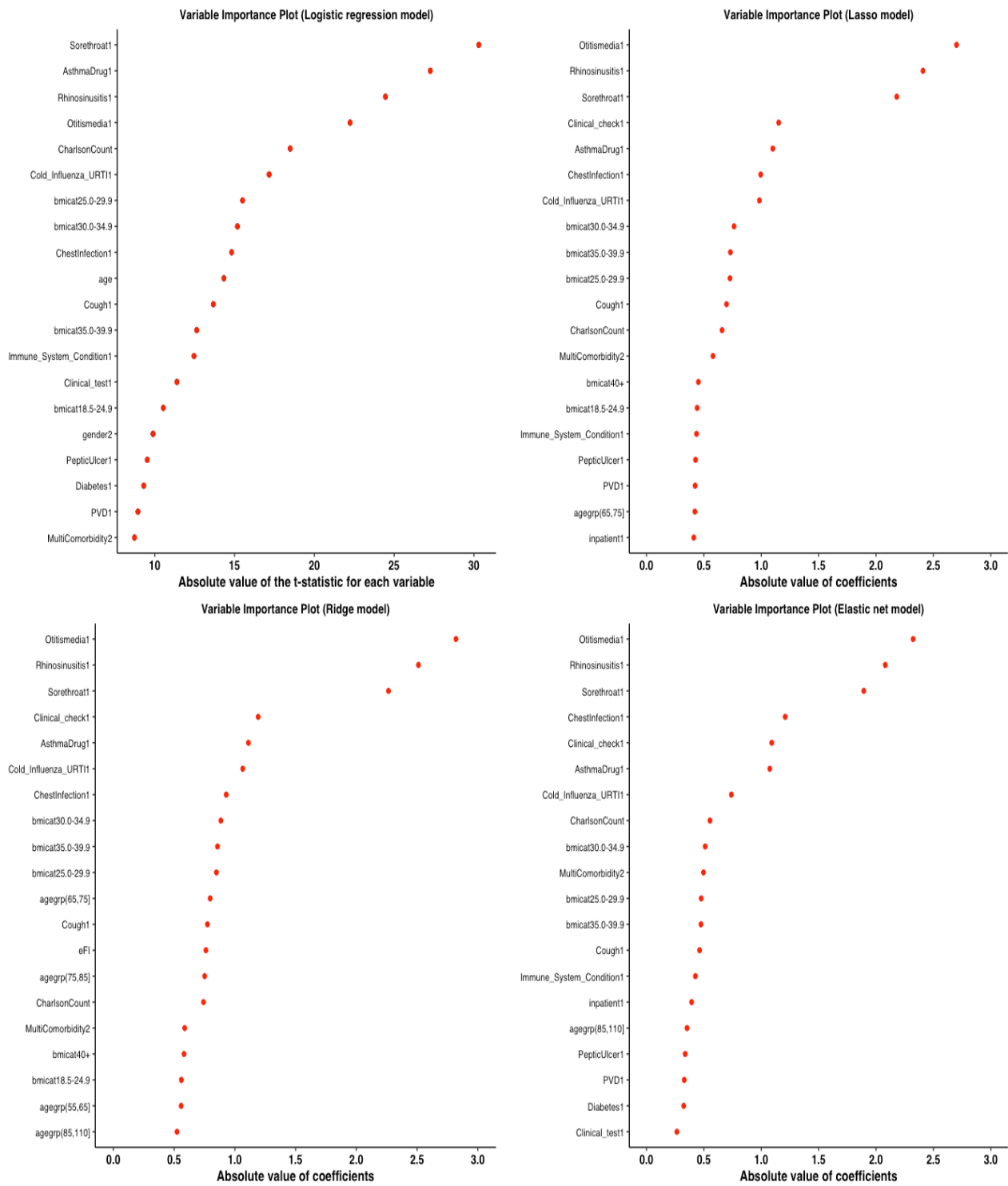


Figure 8.3: Variable importance for full model based on simple logistic regression and penalized regressions

For the elastic net procedure, minimum lambdas and the ones plus their one SE for 100 alphas ranging from 0 to 1 with 0.001 increments were identified using 10-fold cross-validation. The results are summarised in Table A 12 with corresponding misclassification errors reported. Alpha values against misclassification errors are plotted in Figure A 3.

The top 20 important variables were assigned with scores and summarised across random forest and penalized regression algorithms in Table A 11 and Table A 14. Then, the top 15 important variables for three models were presented in Table 8.6. The top four important variables were: chest infection, BMI category, Charlson comorbidity count and immune system condition. The rest as shown to be important by two algorithms were highlighted as yellow in Table 8.6.

Table 8.6: Top 15 variables as selected by simple logistic regression, penalized regression and random forest (full model)

Logistic regression^a	Penalized regression^b	Random forest^c
Sore throat	Otitis media	Age
Asthma drug	Rhinosinusitis	Chest Infection
Rhinosinusitis	Sore throat	eFrailty Index
Otitis media	Clinical check	BMI category
Charlson Count	Asthma drug	Season
Cold/ Influenza/ URTI	Chest infection	Smoking Status
BMI category	Cold/ Influenza/ URTI	Charlson Score
Chest infection	BMI category	Age group
Age	Charlson Count	Charlson Count
Cough	Cough	Antibiotic Ever
Immune system Condition	Multi-comorbidity	eFrailty category
Clinical test	Age group	Gender
Gender	Immune system condition	Antibiotic
Peptic Ulcer	Peptic Ulcer	Immune system condition
Diabetes	eFrailty index	Multi-comorbidity

^a ranking determined by strongest to weakest absolute values of t statistics of simple logistic regression model

^b ranking determined by strongest to weakest absolute values of penalized regression coefficients

^c ranking determined by largest to smallest mean decrease in Gini

Red: Top risk factors for three models

Yellow: Top risk factors in two algorithms

Green: Top risk factors in one algorithm

8.1.3 Model development and performance

Model specification started with comparison of 15 machine learning approaches using the full dataset with 10-fold cross validation through the Classification Learner App in Matlab (The MathWorks, 2020). Comparison results are presented in Table A 15 and Table A 16.

Apart from Gaussian Naïve Bayes, Kernel Naïve Bayes and RUSBoost Tree algorithms, the remaining models reached specificity above 90%. Medium Gaussian SVM has shown to have the highest accuracy (87.6%), and simple logistic regression has best overall discrimination performance as measured by AUC (0.84). Three decision tree models including coarse tree, fine tree and medium tree have similar predictive accuracy (87.0%, 87.2% and 87.1% respectively). Fine tree performed better in terms of AUC and sensitivity compared to the other two decision tree models but with slightly lower specificity. Given that the comparison results between coarse decision tree and medium decision tree are comparable, CART with 5 to 7 levels of split together with simple logistic regression were selected for model specification. The tree approach was considered to have greater clinical utility since decision trees are generally more easily interpreted by a wider audience.

Meanwhile, the simple logistic regression model provides complementary information on the direction and size of predictor effects. Model performance was evaluated by both internal validation and temporal validation. The whole data set was randomly split into 80% for model development and 20% for the internal. A subset of data from 2014 to 2017 was deployed for temporal validation.

Based on variable selection results, variables for CART modelling included: age group, chest infection, Charlson comorbidity count, frailty category, season, smoking status, BMI category, gender and immune system conditions and antibiotic prescription during initial RTI consultations. In the final CART model, four variables including chest infection, antibiotic prescription, age group and Charlson comorbidity count were used for medium tree model with five levels of split. The tuning parameter was selected as shown in Figure A 4.

The CART model is illustrated below in Figure 8.4. Among RTI patients, those who presented with LRTIs labelled as ‘chest infection’ were at higher risk of reconsulting with pneumonia in the subsequent 30 days. Within this patient group, those who did not receive antibiotic prescriptions had the highest probability (70%) of reconsultation with pneumonia. For LRTI patients, even if antibiotic prescriptions were issued by GPs, there were two subgroups of patients who were more likely to reconsult with pneumonia in the following 30 days: LRTI patients aged 85 and above and those age between 76 and 85 with three or more comorbidities measured by Charlson comorbidity index.

Both internal and temporal validation performances as shown in Table 8.7, Figure 8.5 and Figure 8.6 demonstrated that this CART model has low sensitivities (0.25 for internal validation and 0.21 for temporal validation) but high specificities (0.98 for both validations). Comparison statistics of development data and temporal validation data is shown in Table A 17. This gives model discrimination performances being 0.70 and 0.68 for internal and temporal validations respectively. This CART model is not replicable using recent data as shown by the calibration performance of temporal validation as measured by H-L test (p value < 0.05).

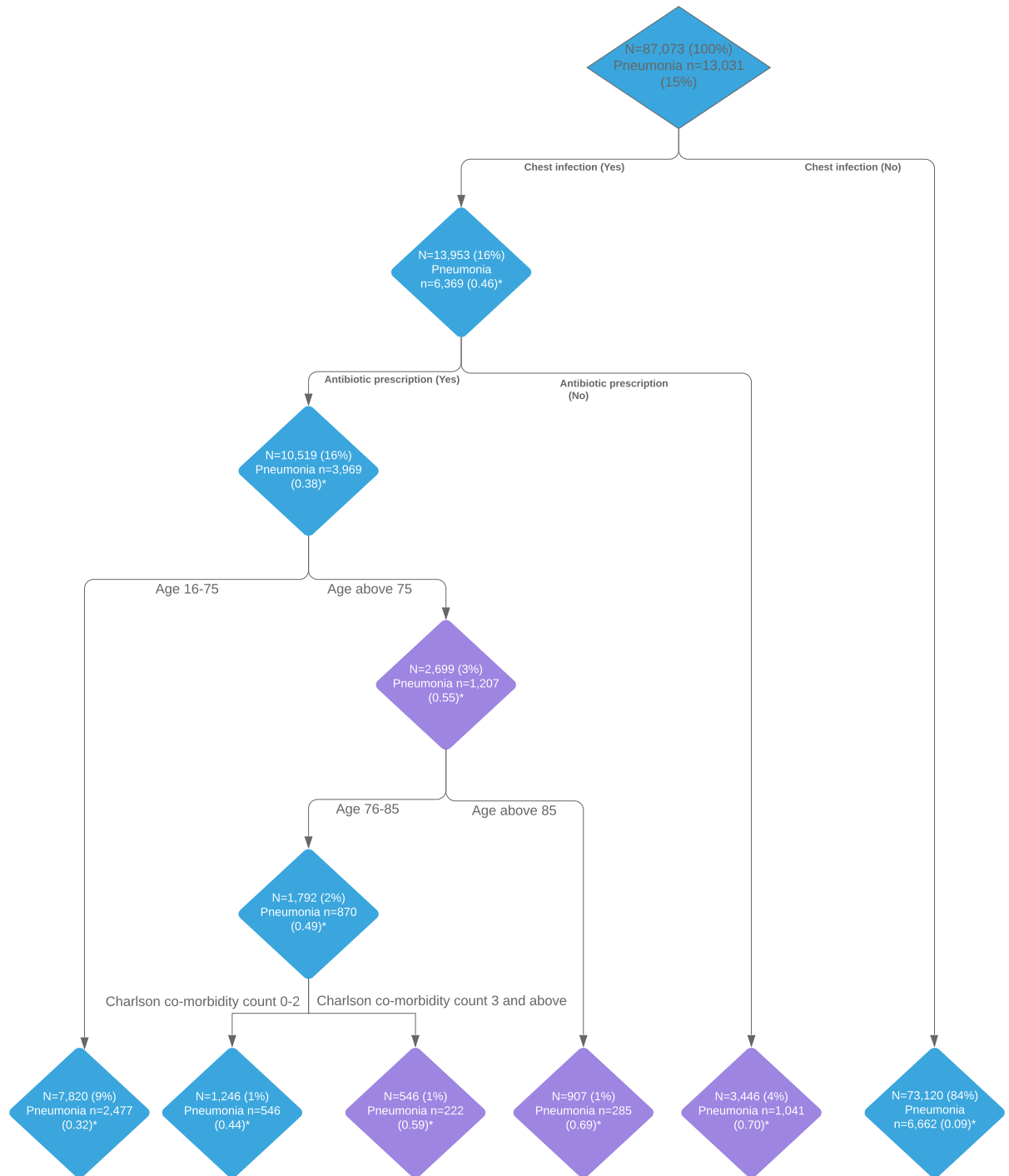


Figure 8.4: Classification tree based on variable selection results for study population (full model)

* Probabilities of developing pneumonia among patients in individual nodes or leaves

Purple: Majority of patients in individual nodes being pneumonia cases

Blue: Majority of patients in individual nodes being non-pneumonia cases

Table 8.7: Classification and regression tree (CART) model performance for internal and temporal validations

	Internal Validation	Temporal Validation
Sensitivity	0.25 (0.23, 0.26)	0.21 (0.20, 0.23)
Specificity	0.98 (0.98, 0.98)	0.98 (0.98, 0.99)
Positive predictive value	0.70 (0.67, 0.72)	0.71 (0.68, 0.74)
Negative predictive value	0.88 (0.88, 0.89)	0.87 (0.87, 0.88)
Positive likelihood ratio	13.16 (11.67, 14.84)	13.59 (11.89, 15.53)
Negative likelihood ratio	0.77 (0.75, 0.78)	0.39 (0.37, 0.41)
AUROC	0.70 (0.69, 0.71)	0.68 (0.67, 0.69)
H-L test	p-value = 1	p-value < 0.01

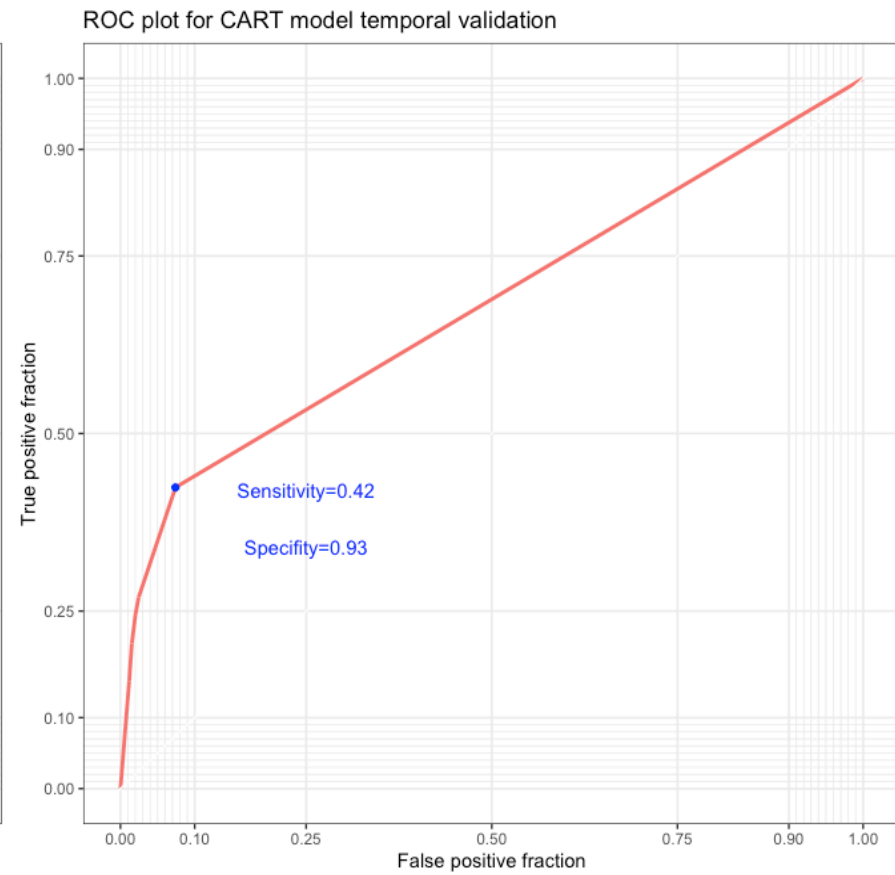
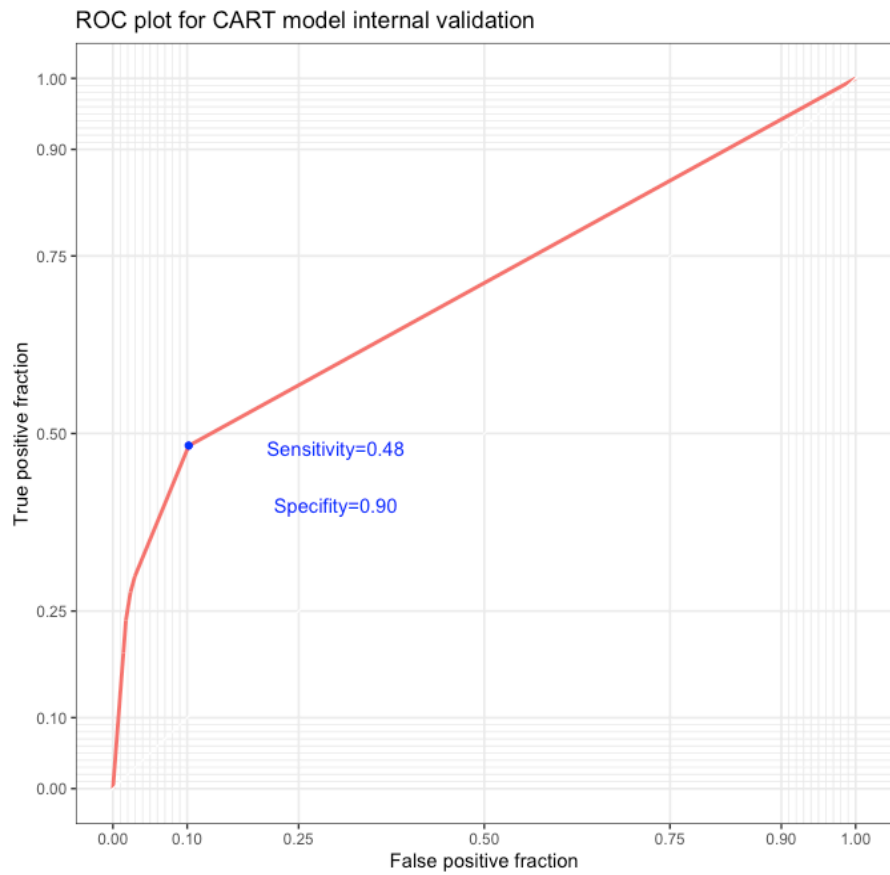


Figure 8.5: Receiver operating characteristic (ROC) curve for internal and temporal validation of CART model (full model)

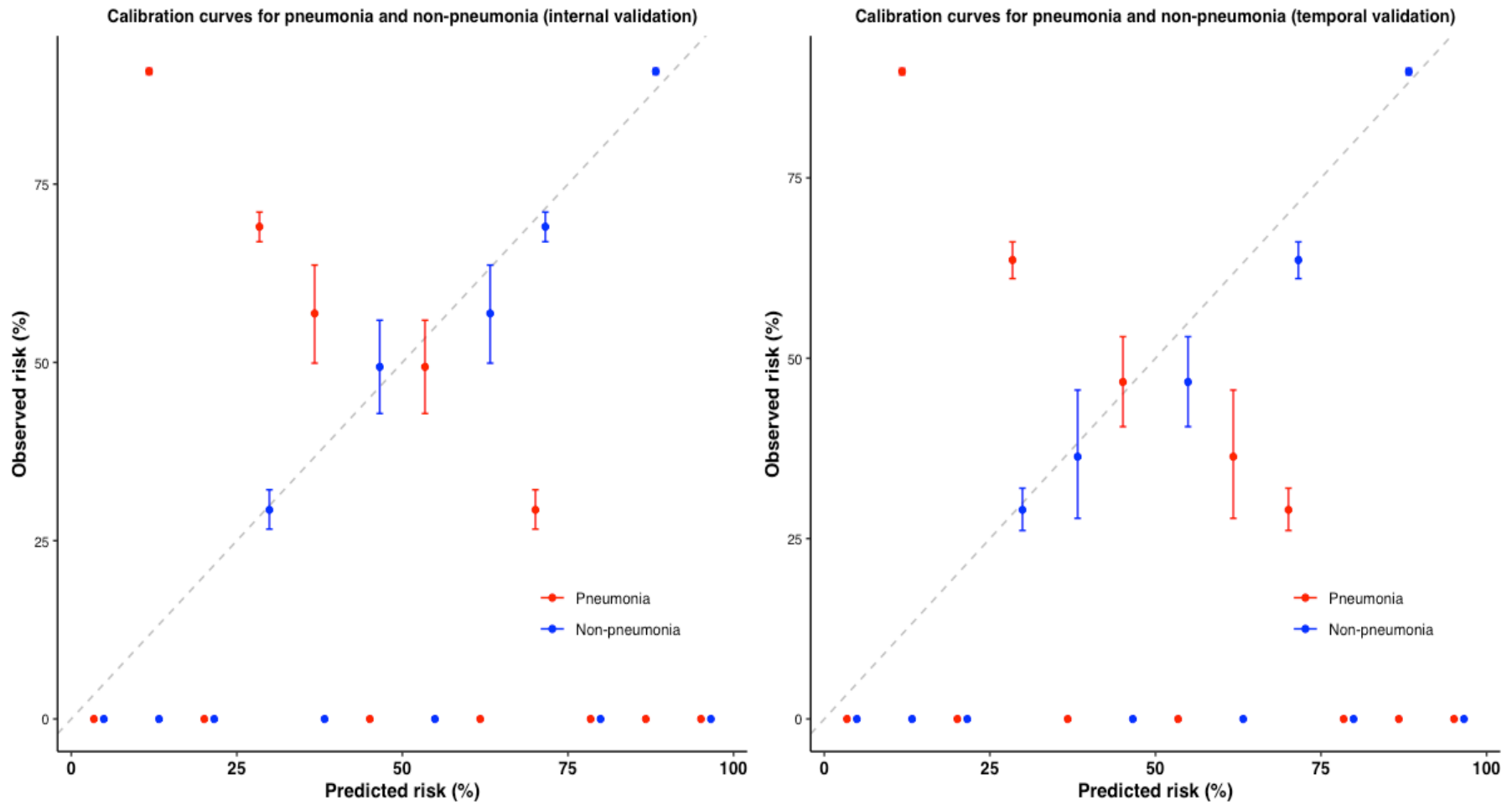


Figure 8.6: Calibration plots for internal and temporal validation of CART model (full model)

The same four predictors for the CART model including chest infection, antibiotic prescription, age group and Charlson co-morbidity count were fitted into simple logistic regression model with their effects quantified show in Table 8.8.

Patients reconsulted with pneumonia within 30 days after previous RTI consultations had 7.08 (6.76 to 7.42) higher odds of presenting with LRTI symptoms during initial visits than those free of pneumonia re-consultations. The odds of RTI patients reconsulting with pneumonia within 30 days increased parallelly with age with the highest odds (OR 7.14, 95% CI (6.48 to 7.68)) found among patients older than 85. A similar trend is noted for Charlson comorbidity counts, that is, pneumonia re-consultations were more likely to be found among RTI patients with multi-comorbidities. Antibiotic prescriptions were shown to be protective for RTI patients from pneumonia re-consultations (OR 0.73, 95% CI (0.70 to 0.76)).

Both internal and temporal validation performances of simple logistic regression model as shown in Table 8.9, Figure 8.7 and Figure 8.8. Similar to the CART model, low sensitivities (0.27 for internal validation and 0.29 for temporal validation) but high specificities (0.97 for both validations) are noticed. Better discrimination performances with statistical differences compared to the CART model are demonstrated with 0.81 and 0.80 for internal and temporal validations respectively. This simple logistic regression model is replicable as shown by the calibration performances measured by H-L test (p value > 0.05 for both internal and temporal validations).

Table 8.8: Simple logistic model using same variables from those for CART (full model)

	Odds ratio	95% CI	p value
Age group			
Age group (16,35] (Ref)			
Age group (35,45]	1.65	(1.51, 1.79)	<0.01
Age group (45,55]	1.66	(1.53, 1.81)	<0.01
Age group (55,65]	2.18	(2.01, 2.36)	<0.01
Age group (65,75]	2.71	(2.5, 2.94)	<0.01
Age group (75,85]	3.91	(3.6, 4.25)	<0.01
Age group (85,110]	7.14	(6.48, 7.86)	<0.01
Charlson comorbidity count			
Charlson Count (0) (Ref)			
Charlson Count (1)	1.28	(1.21, 1.34)	<0.01
Charlson Count (2)	1.87	(1.75, 2.00)	<0.01
Charlson Count (3)	2.37	(2.17, 2.58)	<0.01
Charlson Count (4)	3.18	(2.82, 3.59)	<0.01
Charlson Count (5)	3.52	(2.92, 4.25)	<0.01
Charlson Count (6)	4.86	(3.67, 6.43)	<0.01
Antibiotic prescription (Yes)	0.73	(0.70, 0.76)	<0.01
Chest infection (Yes)	7.08	(6.76, 7.42)	<0.01

Table 8.9: Simple logistic model performance for internal and temporal validations (full model)

	Internal Validation	Temporal Validation
Sensitivity	0.27 (0.26, 0.29)	0.29 (0.27, 0.30)
Specificity	0.97 (0.97, 0.98)	0.97 (0.97, 0.97)
Positive predictive value	0.64 (0.61, 0.67)	0.65 (0.62, 0.67)
Negative predictive value	0.88 (0.88, 0.89)	0.88 (0.88, 0.89)
Positive likelihood ratio	10.09 (9.10, 11.19)	10.01 (9.05, 11.07)
Negative likelihood ratio	0.75 (0.73, 0.77)	0.74 (0.72, 0.75)
AUROC	0.81 (0.81, 0.84) ^a	0.80 (0.79, 0.81) ^a
H-L test	p-value = 0.09	p-value = 0.14

^aDifference in AUROC between CART and simple logistic regression models is significant (p value<0.01)

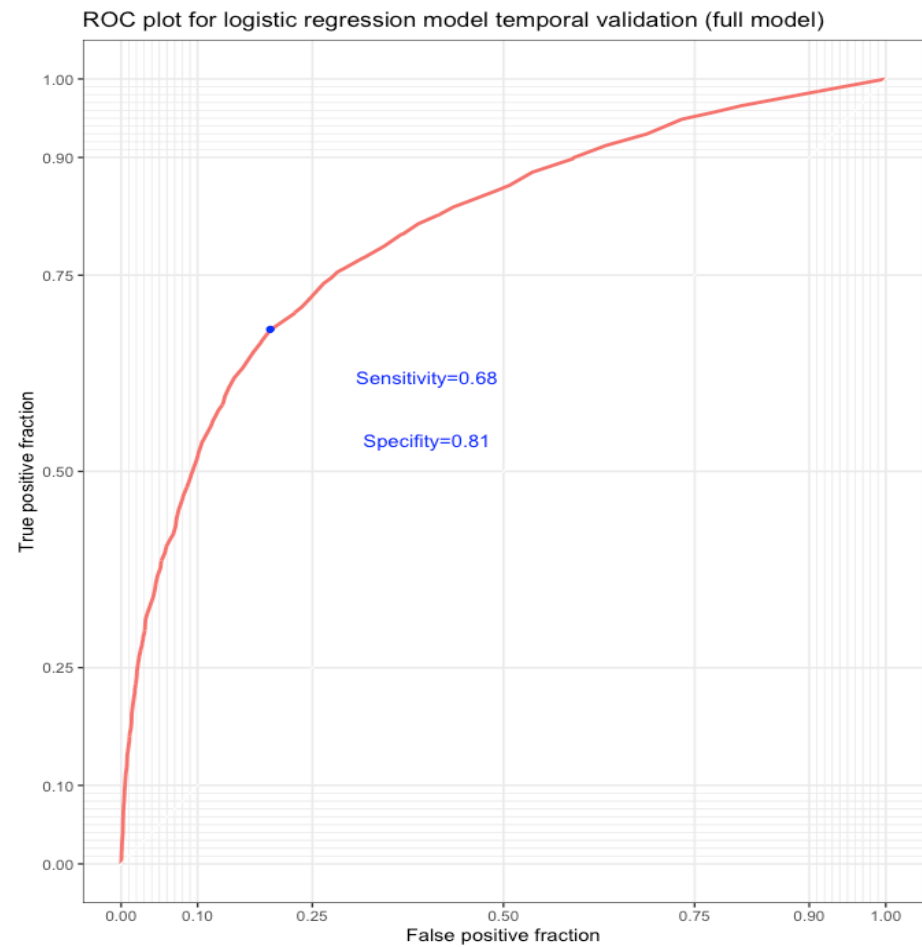
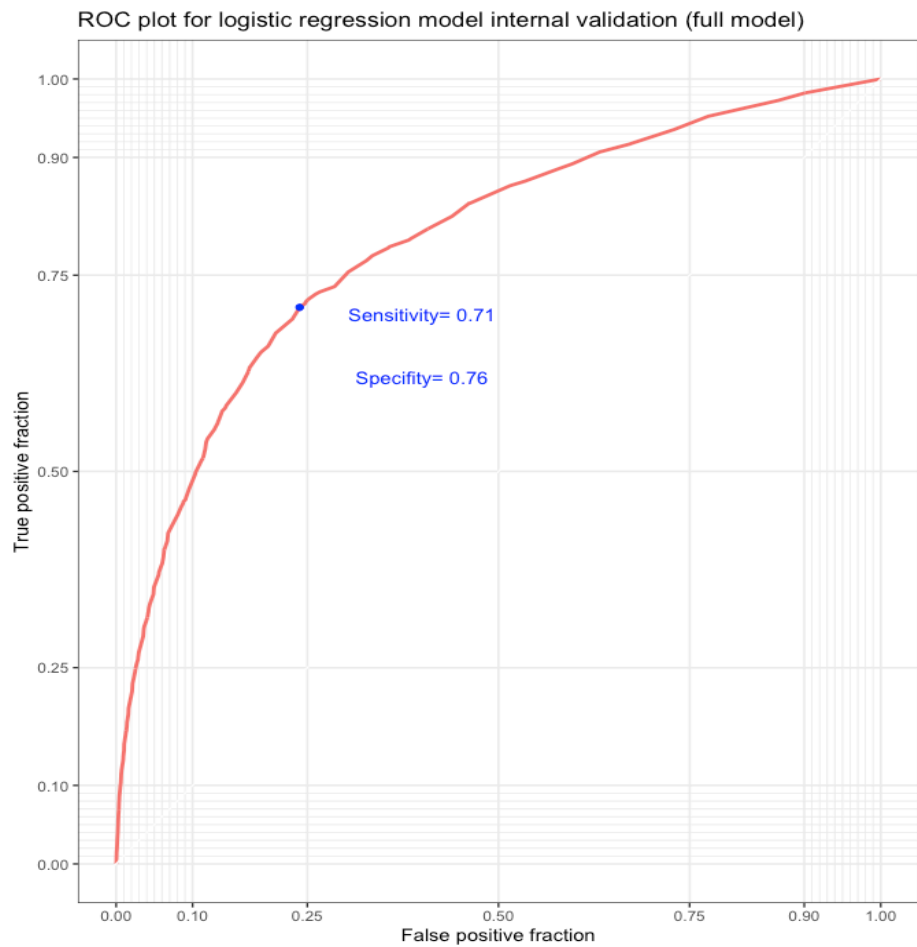


Figure 8.7: ROC curve for internal and temporal validation of simple logistic model (full model)

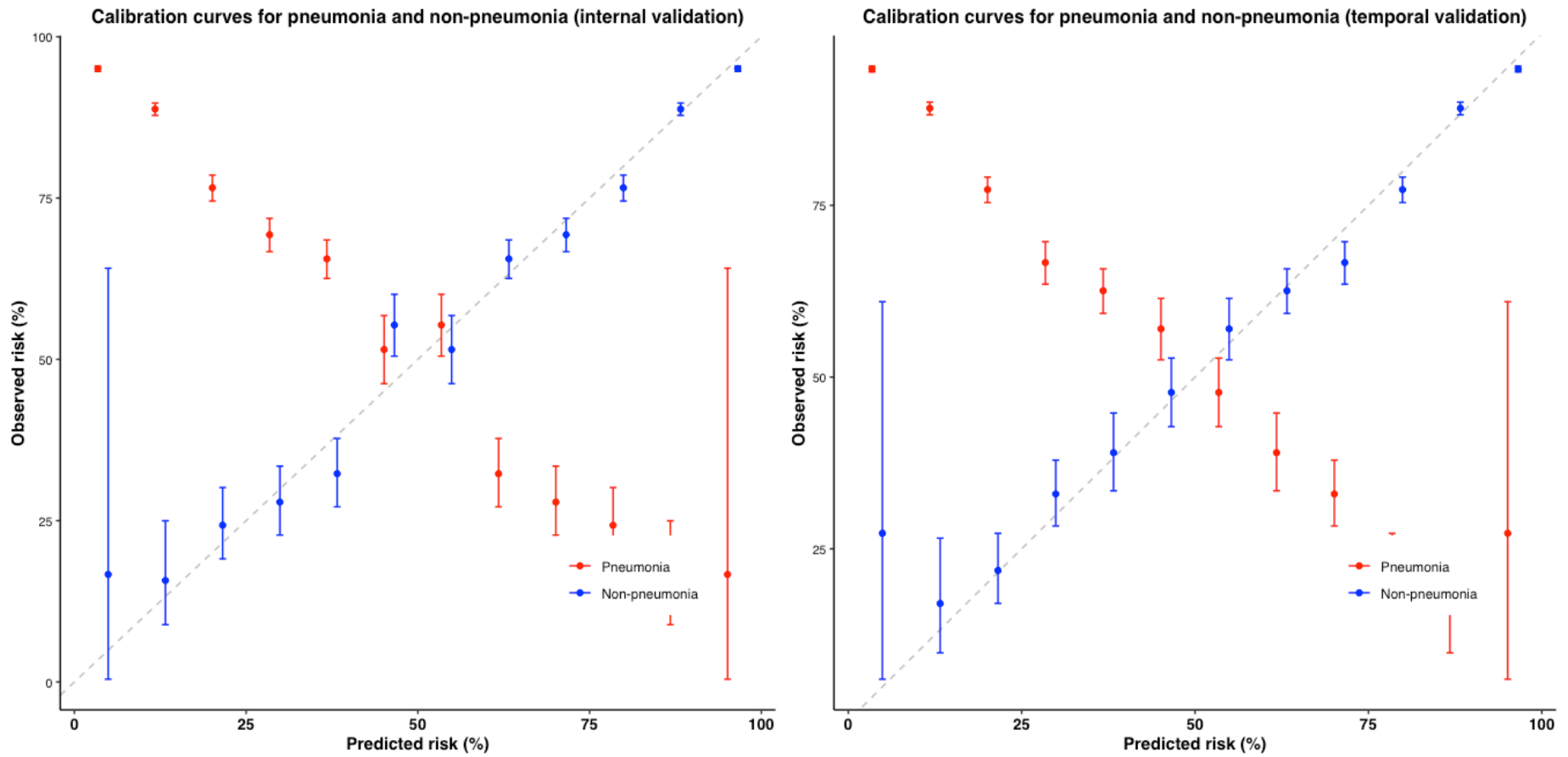


Figure 8.8: Calibration plots for internal and temporal validation of simple logistic regression (full model)

8.1.4 Subgroup analysis

Based on variable selection and CART modelling results, LRTIs labelled as ‘chest infection’ represent a strong predictor for the onset of pneumonia following RTI consultations within 30 days. The same modelling procedures and performance evaluations are applied to patients presented with LRTI symptoms and those with URTIs only.

8.1.4.1 LRTI patients

The top 15 important variables for three models (random forest, penalised regression models and simple logistic regression model after backward variable selection) are presented in Table 8.10 with results in each process reported in Table A 18, Table A 19, Table A 20, Table A 21 and Figure A 5, Figure A 6, Figure A 7, Figure A 8, Figure A 9. The top seven important variables identified by three modelling approaches are: Charlson comorbidity count, eFrailty index, BMI category, antibiotic prescription before pneumonia index date, asthma drug use, immune system conditions and multi-comorbidity. Age, age group, gender, Charlson score, cough and chronic heart disease are ranked as important by two algorithms.

Table 8.10: Top 15 variables as selected by simple logistic regression, penalized regression and random forest for lower respiratory tract infection (LRTI) patients

Logistic regression ^a	Penalized regression ^b	Random forest ^c
Antibiotic ever	Rhinosinusitis	Age
Cough	Clinical check	eFrailty Index
Charlson count	Sore throat	BMI category
Asthma drug	Otitis media	Season
Immune system condition	Antibiotic ever	Antibiotic Ever
Clinical test	Cough	Age group
BMI category	Asthma drug	Charlson Score
Age	BMI category	Smoking Status
Chronic heart disease	Charlson count	Charlson Count
Multi-comorbidity	Cold/ Influenza/ URTI	Immune system condition
Gender	Multi-comorbidity	Antibiotic
PVD	Age group	Asthma drug
Charlson score	Immune system condition	eFrailty category
Peptic Ulcer	eFrailty index	Gender
eFrailty Index	Chronic heart disease	Multi-comorbidity

^a ranking determined by strongest to weakest simple logistic regression coefficients

^b ranking determined by strongest to weakest penalized regression coefficients

^c ranking determined by largest to smallest mean decrease in Gini

Red: Top risk factors for three models

Yellow: Top risk factors in two algorithms

Green: Top risk factors in one algorithm

Based on variable selection results, variables for CART modelling included: age group, Charlson comorbidity count, BMI category, asthma drug use, immune system conditions, multi-comorbidity and antibiotic prescription during initial RTI consultations. In the final CART model, five variables including antibiotic prescription, age group, Charlson co-morbidity count, asthma drug use and immune system conditions were used for medium tree model with five levels of split. The tuning parameter was selected as shown in Figure A 10.

The CART model is illustrated below in Figure 8.9. Among LRTI patients, those who did not receive antibiotic prescriptions were at higher risk of reconsulting with pneumonia in the subsequent 30 days. Within this patient group, patients who were aged between 16 and 65 and were not under asthma drug treatment nor immunosuppressant, the re-consultation risk is estimated to be 65%. For LRTI patients, if antibiotics were not issued to patients older than 65, the risk of pneumonia re-consultation is about 0.79. On the other hand, even if antibiotic prescriptions were issued by GPs, there were two subgroups of patients who were more likely to reconsult with pneumonia in the following 30 days: patients aged 85 and above and those age between 76 and 85 with two or more comorbidities measured by Charlson comorbidity index.

Both internal and temporal validation performances as shown in Table 8.11, Figure 8.10 and Figure 8.11 demonstrated that this CART model for LRTI patients has acceptable sensitivities (0.56 for internal validation and 0.57 for temporal validation) and specificities (0.78 for internal validations and 0.76 for temporal validation). Comparison statistics of development data and temporal validation data is shown Table A 22. This gives model discrimination performances being 0.69 for both internal and temporal validations. This CART model is not replicable using recent data as shown by the calibration performance of temporal validation as measured by H-L test (p value < 0.05).

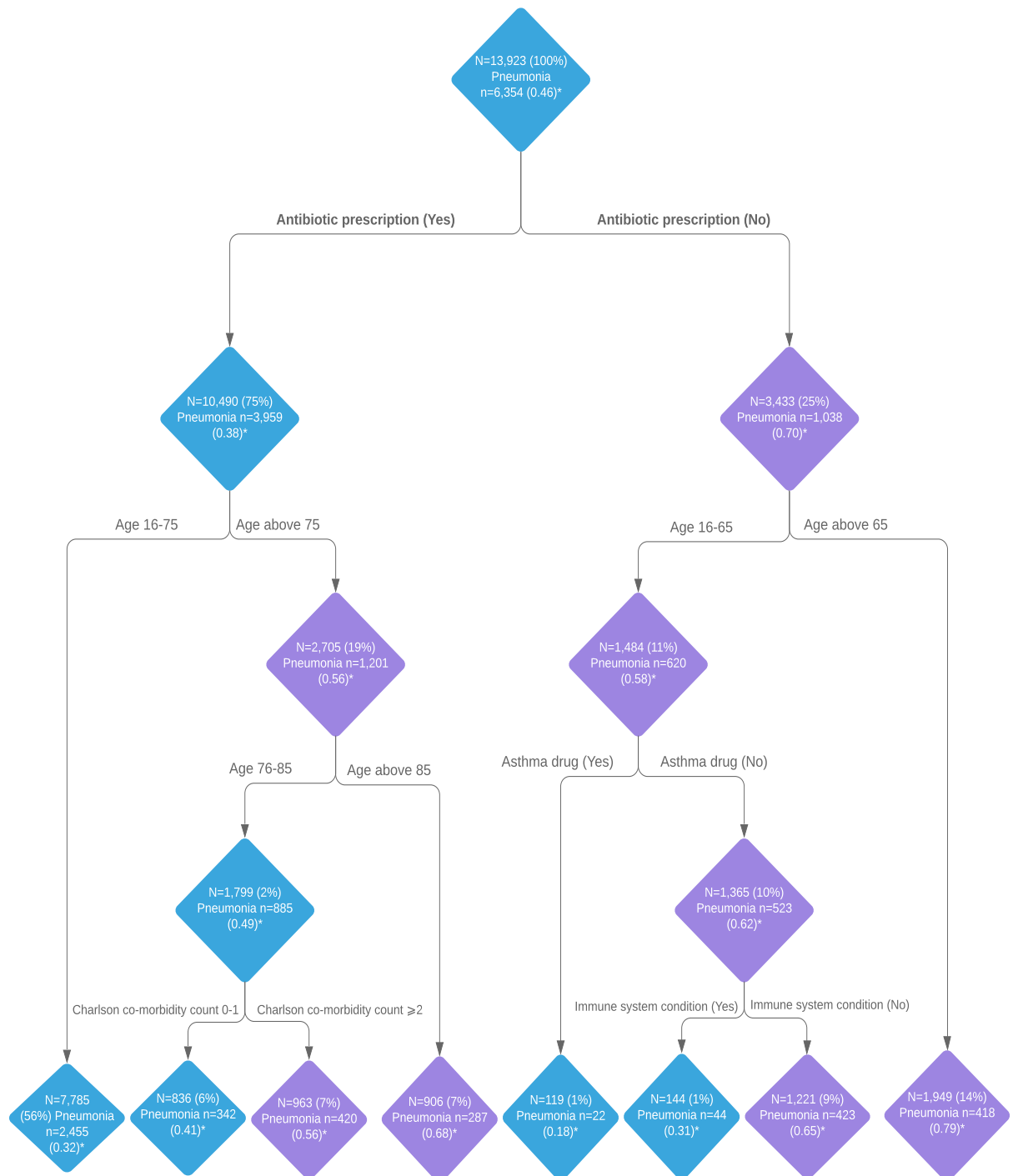


Figure 8.9: CART model for patients presented with LRTIs

* Probabilities of developing pneumonia among patients in individual nodes or leaves

Purple: Majority of patients in individual nodes being pneumonia cases

Blue: Majority of patients in individual nodes being non-pneumonia cases

Table 8.11: CART model performance for internal and temporal validations (LRTI model)

	Internal Validation	Temporal Validation
Sensitivity	0.56 (0.53, 0.58)	0.57 (0.54, 0.59)
Specificity	0.78 (0.76, 0.80)	0.76 (0.74, 0.78)
Positive predictive value	0.68 (0.66, 0.71)	0.71 (0.68, 0.74)
Negative predictive value	0.68 (0.66, 0.71)	0.63 (0.60, 0.65)
Positive likelihood ratio	2.55 (2.32, 2.81)	2.36 (2.12, 2.62)
Negative likelihood ratio	0.57 (0.54, 0.60)	0.57 (0.53, 0.61)
AUROC	0.69 (0.67, 0.70)	0.69 (0.67, 0.71)
H-L test	p value: 0.99	p value: 0.002

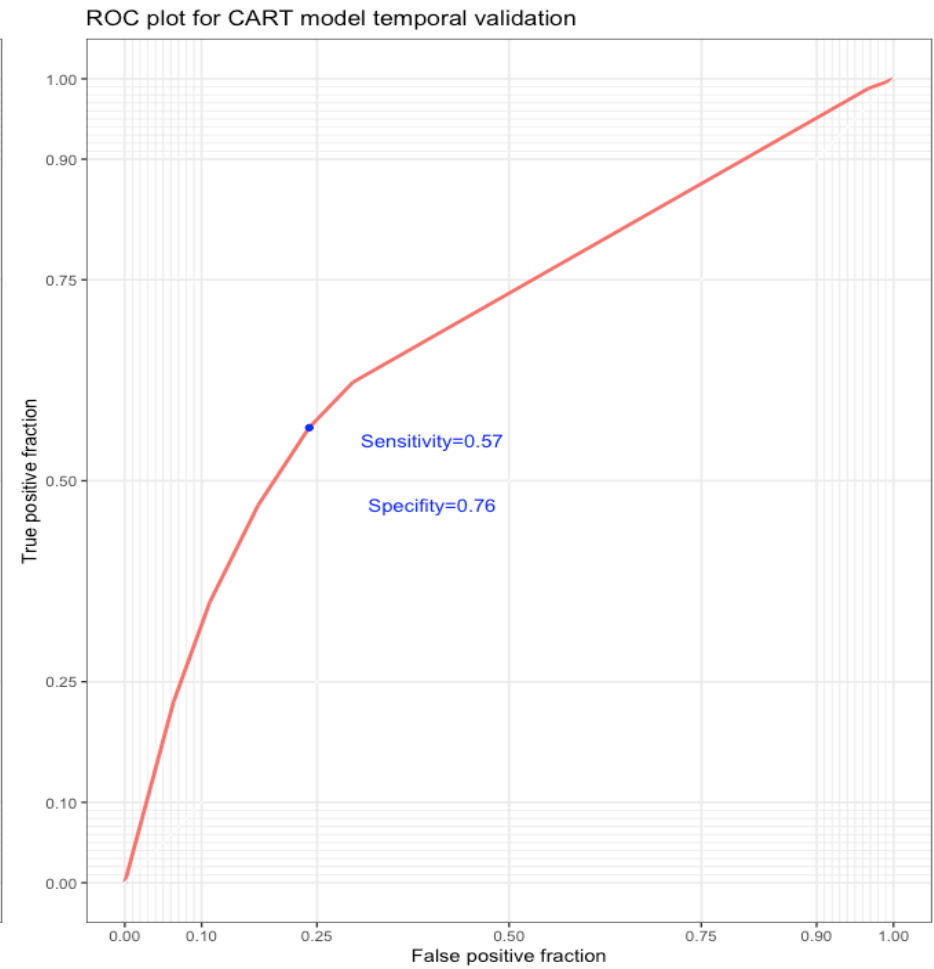
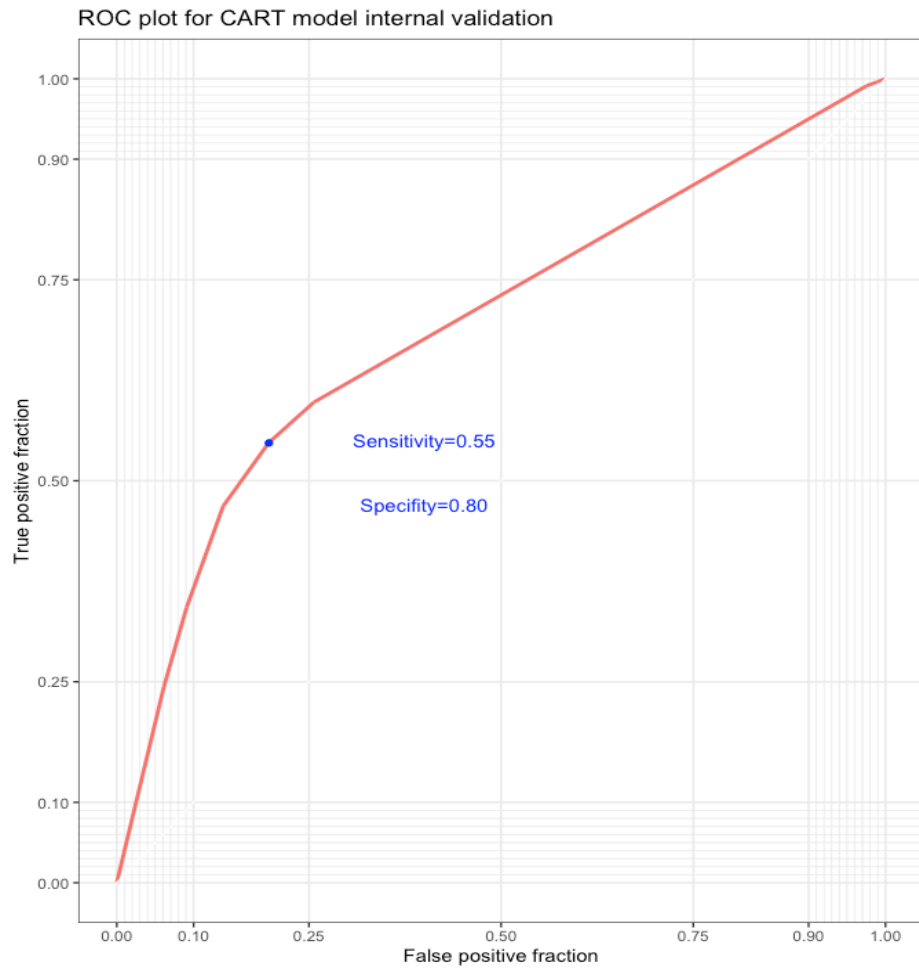


Figure 8.10: ROC curve for internal and temporal validation of CART model (LRTI model)

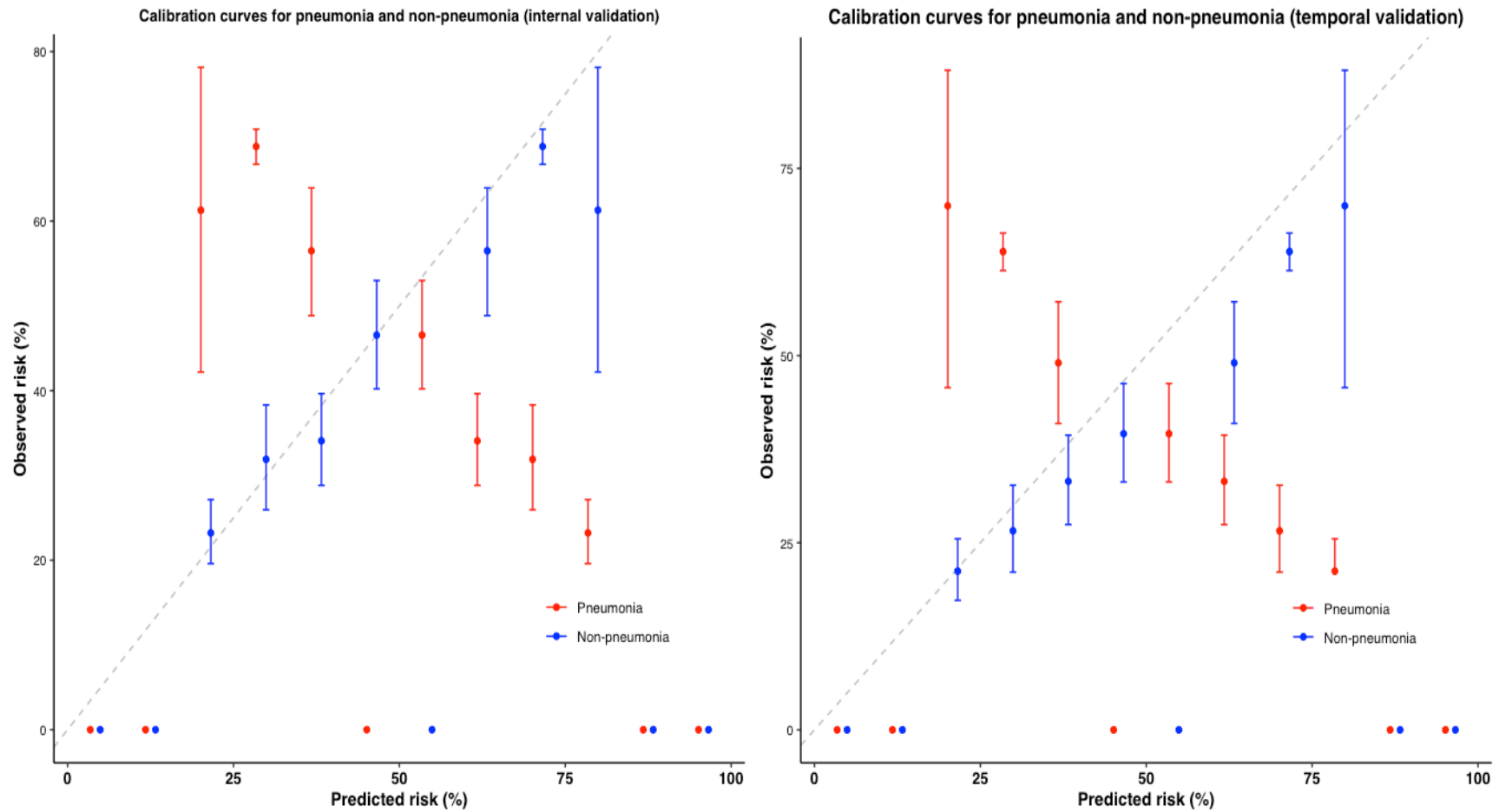


Figure 8.11: Calibration plots for internal and temporal validation of CART (LRTI model)

The same five predictors for CART model including, antibiotic prescription, age group, Charlson co-morbidity count, asthma drug use and immune system conditions were fitted into simple logistic regression model with their effects quantified show in Table 8.12.

Patients reconsulted with pneumonia within 30 days after previous LRTI consultations were 0.29 (0.26 to 0.31) times less likely being prescribed with antibiotics during initial visits than those free of pneumonia re-consultations. The likelihoods of LRTI patients reconsulted with pneumonia within 30 days increased parallelly with age with the highest odds (OR 5.22, 95% CI (4.39 to 6.21)) found among patients older than 85. Similar trend is noted for Charlson comorbidity counts, that is, pneumonia re-consultations were more likely to be found among LRTI patients with multi-comorbidities. Both asthma drug use and immune system conditions have shown to be protective for LRTI patients from pneumonia re-consultations (OR 0.35, 95% CI (0.30 to 0.40) and OR 0.47, 95% CI (0.43 to 0.52) respectively).

Both internal and temporal validation performances of simple logistic regression model as shown in Table 8.13, Figure 8.12 and Figure 8.13. Similar to the CART model, acceptable sensitivities (0.57 for internal validation and 0.69 for temporal validation) and specificities (0.78 for internal validations and 0.69 for temporal validation) are noticed. Discrimination performances with statistical differences compared to the CART model are demonstrated with 0.75 and 0.45 for internal and temporal validations respectively. CART model outperformed simple logistic regression model in temporal validation. This simple logistic regression model is not replicable as shown by the calibration performances measured by H-L test (p value < 0.05 for both internal and temporal validations).

Table 8.12: Simple logistic model using same variables from those for CART (LRTI model)

	Odds ratio	95% CI	p value
Age group			
Age group (16,35] (Ref)			
Age group (35,45]	1.42	(1.21, 1.66)	<0.01
Age group (45,55]	1.33	(1.14, 1.55)	<0.01
Age group (55,65]	1.6	(1.38, 1.85)	<0.01
Age group (65,75]	1.88	(1.62, 2.18)	<0.01
Age group (75,85]	2.82	(2.43, 3.28)	<0.01
Age group (85,110]	5.22	(4.39, 6.21)	<0.01
Charlson comorbidity count			
Charlson Count (0) (Ref)			
Charlson Count (1)	1.49	(1.36, 1.63)	<0.01
Charlson Count (2)	2.09	(1.85, 2.36)	<0.01
Charlson Count (3)	2.64	(2.26, 3.09)	<0.01
Charlson Count (4)	3.86	(3.10, 4.82)	<0.01
Charlson Count (5)	3.64	(2.68, 5.00)	<0.01
Charlson Count (6)	7.69	(4.35, 14.47)	<0.01
Asthma drug use (Yes)	0.35	(0.30, 0.40)	<0.01
Antibiotic prescription (Yes)	0.29	(0.26, 0.31)	<0.01
Immune system condition (Yes)	0.47	(0.43, 0.52)	<0.01

Table 8.13: Simple logistic model performance for internal and temporal validations (LRTI model)

	Internal Validation	Temporal Validation
Sensitivity	0.57 (0.54, 0.59)	0.69 (0.67, 0.71)
Specificity	0.78 (0.76, 0.79)	0.69 (0.66, 0.71)
Positive predictive value	0.68 (0.65, 0.70)	0.70 (0.68, 0.72)
Negative predictive value	0.68 (0.66, 0.70)	0.68 (0.65, 0.70)
Positive likelihood ratio	2.53 (2.30, 2.78)	2.23 (2.04, 2.43)
Negative likelihood ratio	0.56 (0.53, 0.59)	0.97 (0.97, 0.98)
AUROC	0.75 (0.73, 0.77) ^a	0.45 (0.41, 0.49) ^a
H-L test	p-value = 0.007	p-value = 0.002

^a Difference in AUROC between CART and simple logistic regression models is significant (p value<0.01).

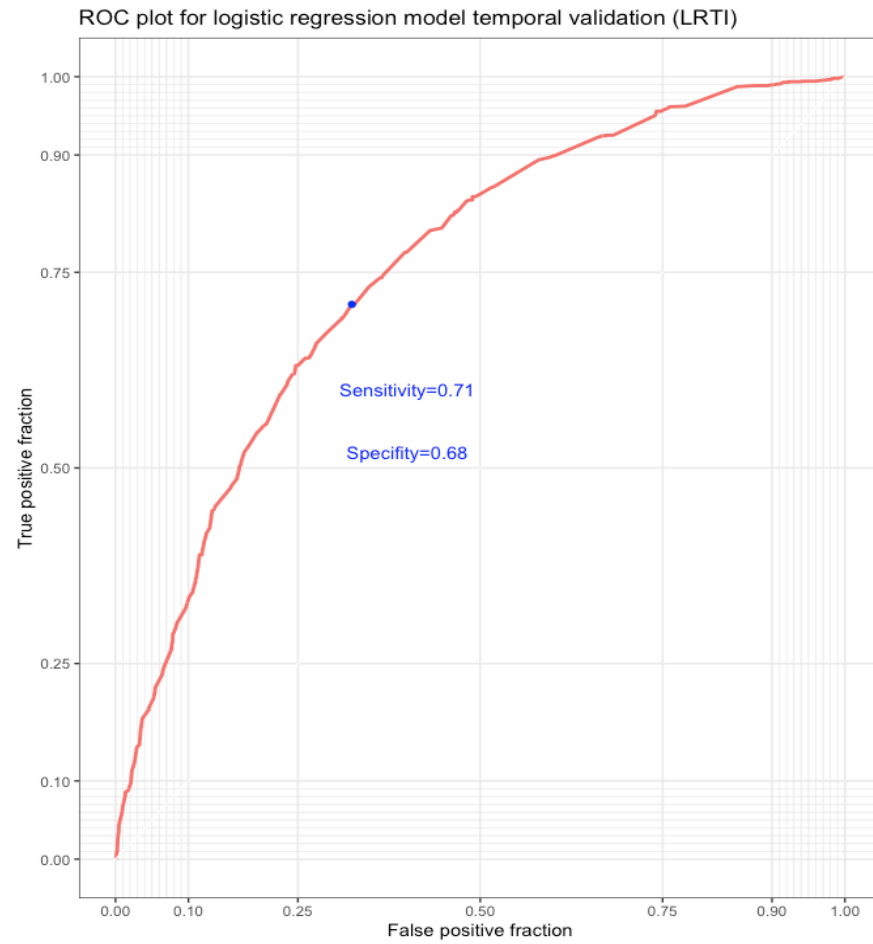
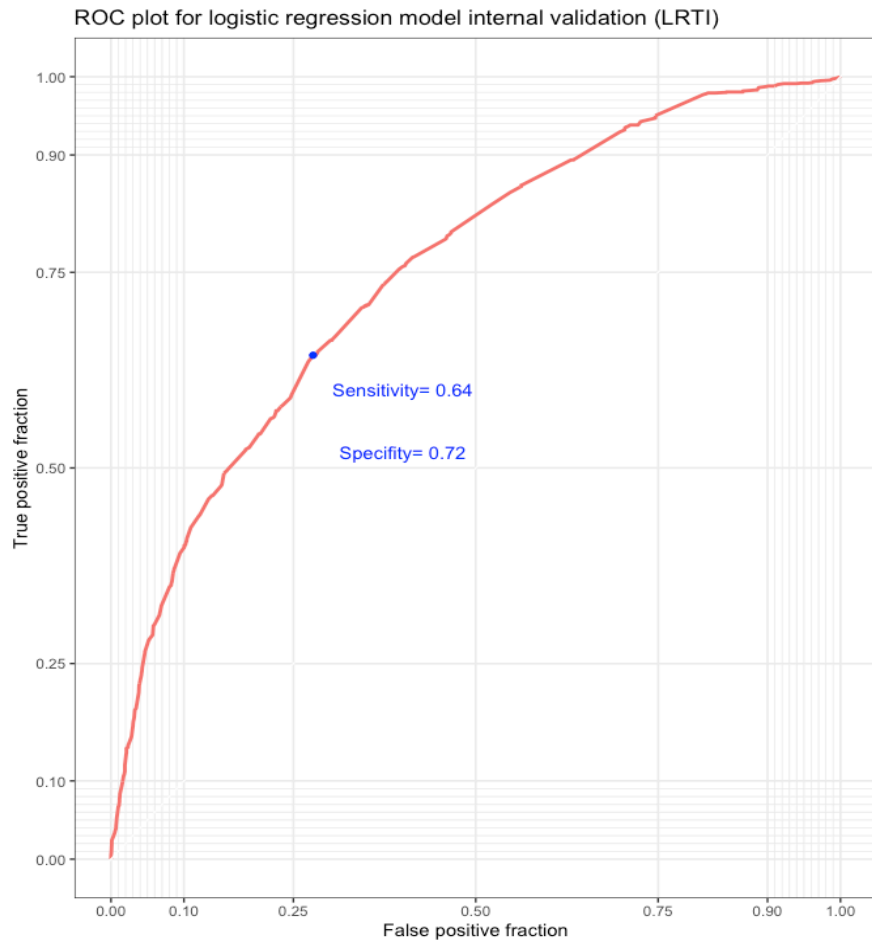


Figure 8.12: ROC curve for internal and temporal validation of simple logistic model (LRTI model)

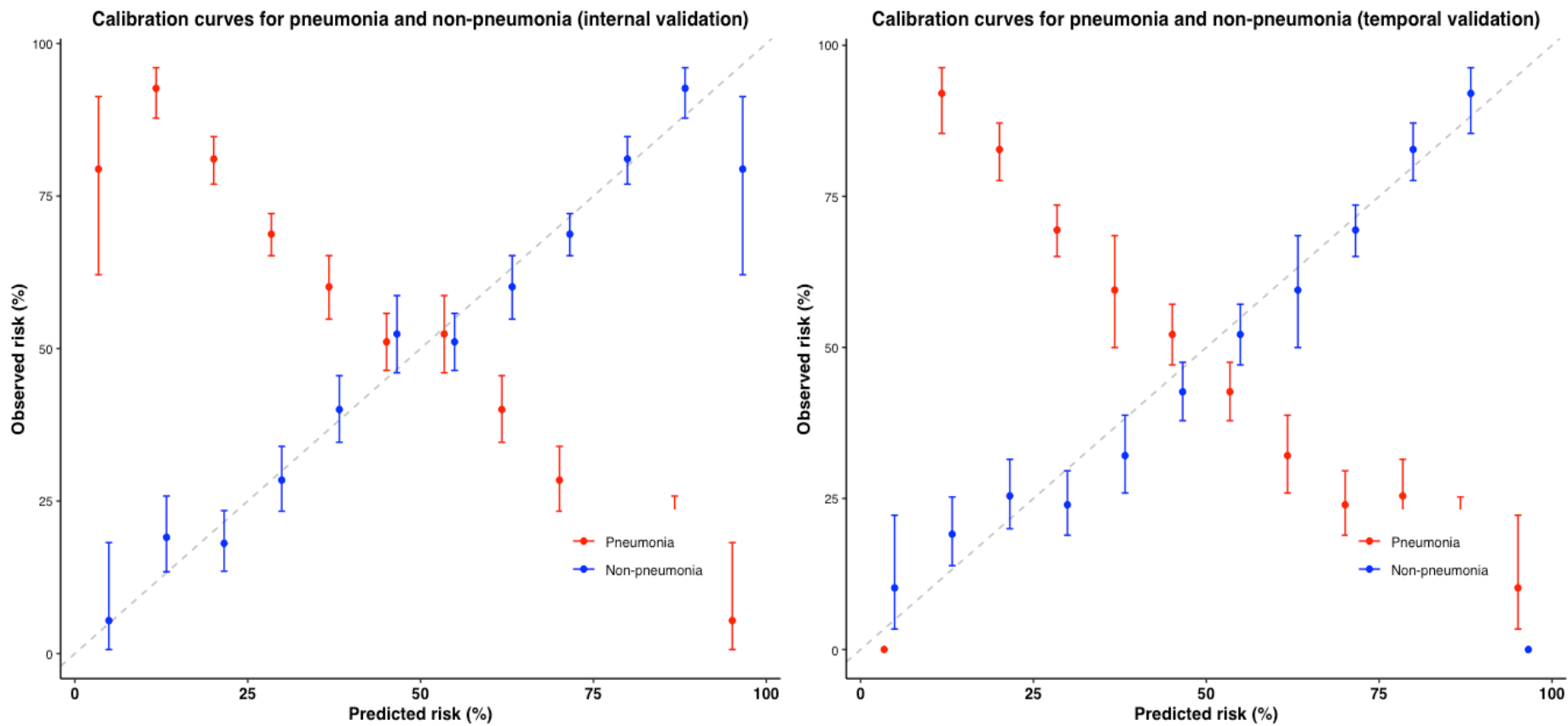


Figure 8.13: Calibration plots for internal and temporal validation of simple logistic regression (LRTI model)

8.1.4.2 URTI patients

The top 10 important variables for three models (random forest, penalised regression models and simple logistic regression model after backward variable selection) are presented in Table 8.14 with results in each process reported in Table A 23, Table A 24, Table A 25, Table A 26 and Figure A 11, Figure A 12, Figure A 13, Figure A 14, Figure A 15. Given that only nine variables are presented with non-zero coefficients by lasso model in Table A 25, number of candidate variables for URTI model is set at 10.

The top two important variables identified by three modelling approaches are: age and Charlson comorbidity count. Age group, eFrailty index, BMI category, Charlson score, asthma drug use, cough, sore throat, rhinosinusitis and otitis media are ranked as important by two algorithms.

Based on variable selection results, variables for CART modelling included: age group, Charlson comorbidity count, BMI category, asthma drug use, eFrailty index, cough, sore throat, rhinosinusitis and otitis media. In the final CART model, four variables age group, Charlson co-morbidity count, asthma drug use and BMI category were used for medium tree model with seven levels of split. Any fewer number of split level than seven only yield a root node. The tuning parameter was selected as shown in Figure A 16.

Table 8.14: Top 10 variables as selected by simple logistic regression, penalized regression and random forest for upper respiratory tract infection (URTI) model

Logistic regression	Penalized regression	Random forest
Sore throat	Asthma drug	Age
Asthma drug	Age group	eFrailty Index
Rhinosinusitis	Cough	BMI category
Otitis media	Sore throat	Season
Charlson Count	Rhinosinusitis	Smoking Status
BMI category	eFrailty index	Charlson Score
Age	Otitis media	Age group
Cough	Charlson Count	Charlson Count
Immune system condition	Charlson Score	Gender
Clinical test	Age	eFrailty category

^a ranking determined by strongest to weakest simple logistic regression coefficients

^b ranking determined by strongest to weakest penalized regression coefficients

^c ranking determined by largest to smallest mean decrease in Gini

Red: Top risk factors for three models

Yellow: Top risk factors in two algorithms

Green: Top risk factors in one algorithm

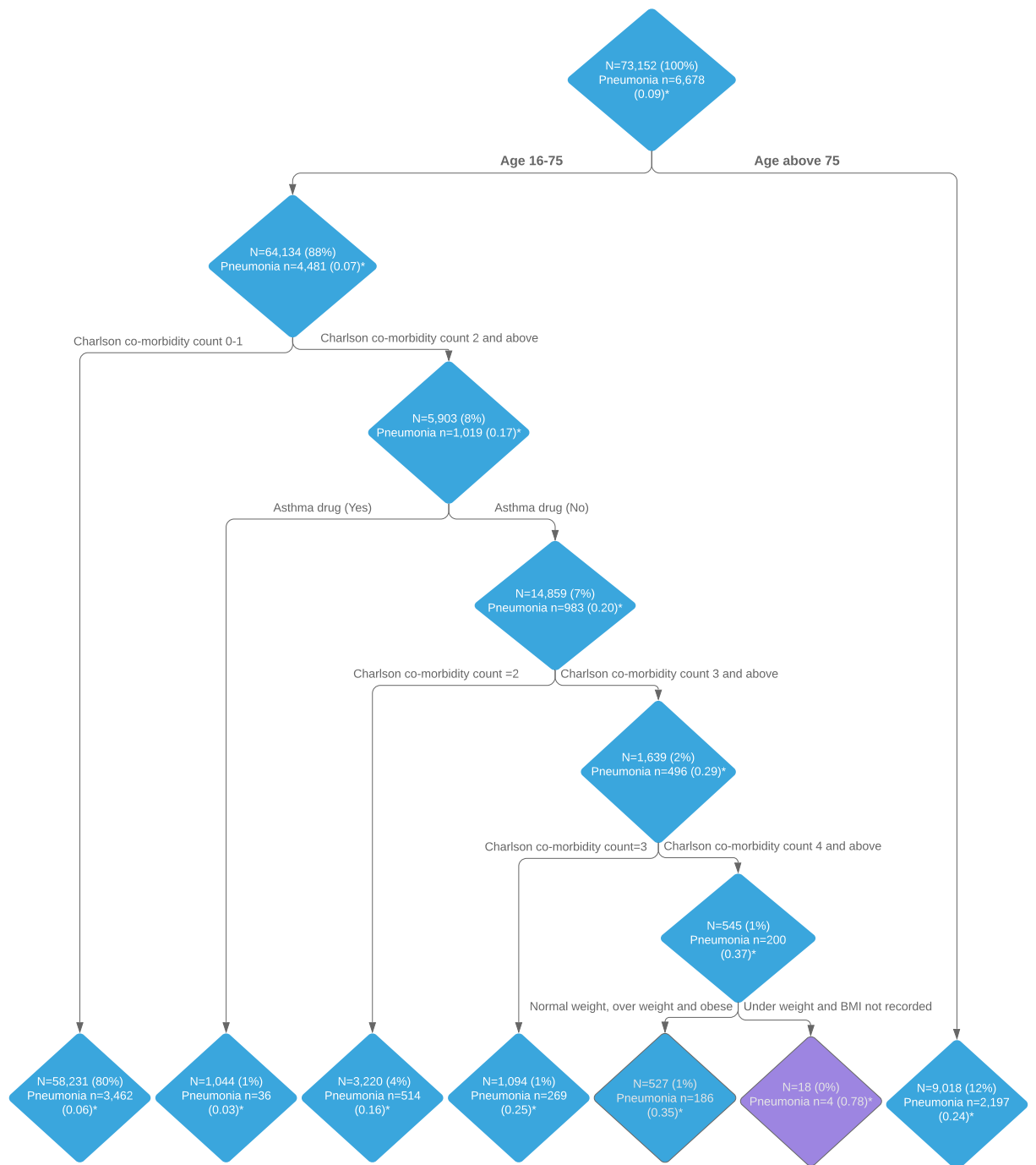


Figure 8.14: CART model for patients presented with URTIs

* Probabilities of developing pneumonia among patients in individual nodes or leaves

Purple: Majority of patients in individual nodes being pneumonia cases

Blue: Majority of patients in individual nodes being non-pneumonia cases

The CART model is illustrated above in Figure 8.14. Among URTI patients, only one subgroup of patients presented with higher risk (0.78) of pneumonia re-consultation: those aged between 16 and 75 without asthma drug treatment but diagnosed with more than three comorbidities, who were under weight or whose BMI information was not documented.

Both internal and temporal validation performances as shown in Table 8.15, Figure 8.15 and Figure 8.16 demonstrated that this CART model for URTI patients has extremely low sensitivity (0 for both internal and temporal validations) but nearly 100% specificity for both validation evaluations. Comparison statistics of development data and temporal validation data is shown Table A 27. This gives model discrimination performances being 0.66 for internal validation 0.68 for temporal validation. This CART model is not replicable using recent data as shown by the calibration performance of temporal validation as measured by H-L test (p value < 0.05).

Table 8.15: CART model performance for internal and temporal validations (URTI model)

	Internal Validation	Temporal Validation
Sensitivity	0.00 (0.00, 0.00)	0.00 (0.00, 0.01)
Specificity	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
Positive predictive value	0.50 (0.01, 0.99)	0.71 (0.29, 0.96)
Negative predictive value	0.91 (0.90, 0.91)	0.90 (0.89, 0.90)
Positive likelihood ratio	9.96 (0.62, 159.1)	21.88 (4.25, 112.72)
Negative likelihood ratio	1.00 (1.00, 1.00)	0.99 (0.98, 0.99)
AUROC	0.66 (0.65, 0.67)	0.68 (0.67, 0.69)
H-L test	p-value = 0.29	p-value = 0.01

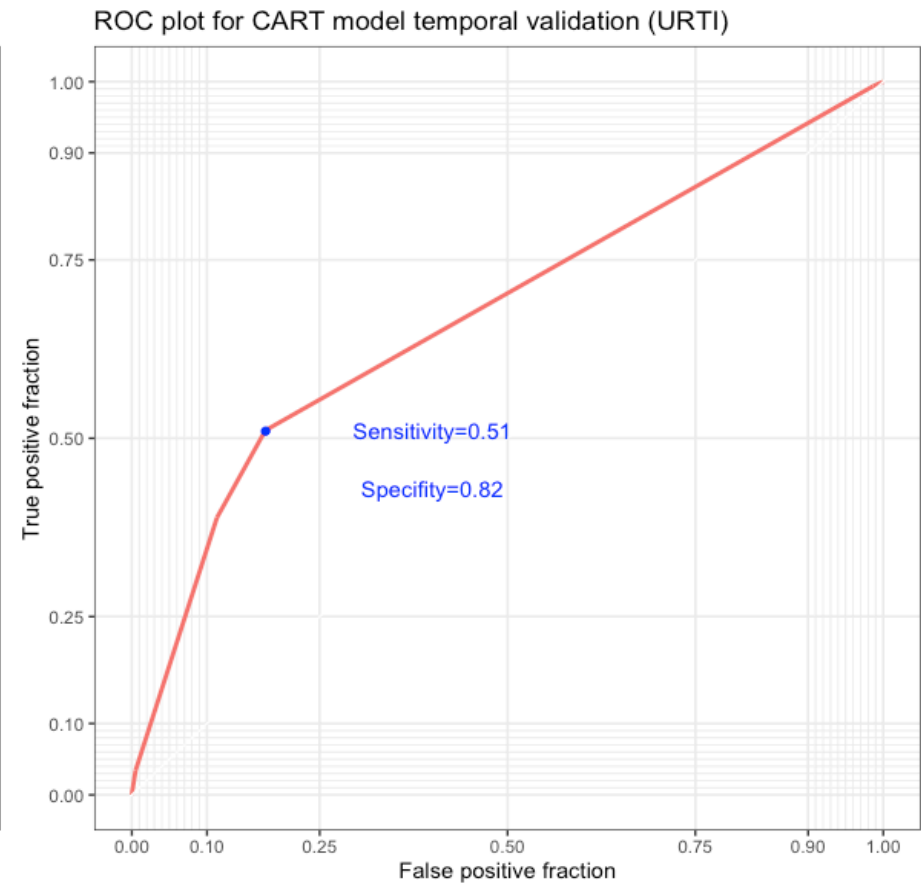
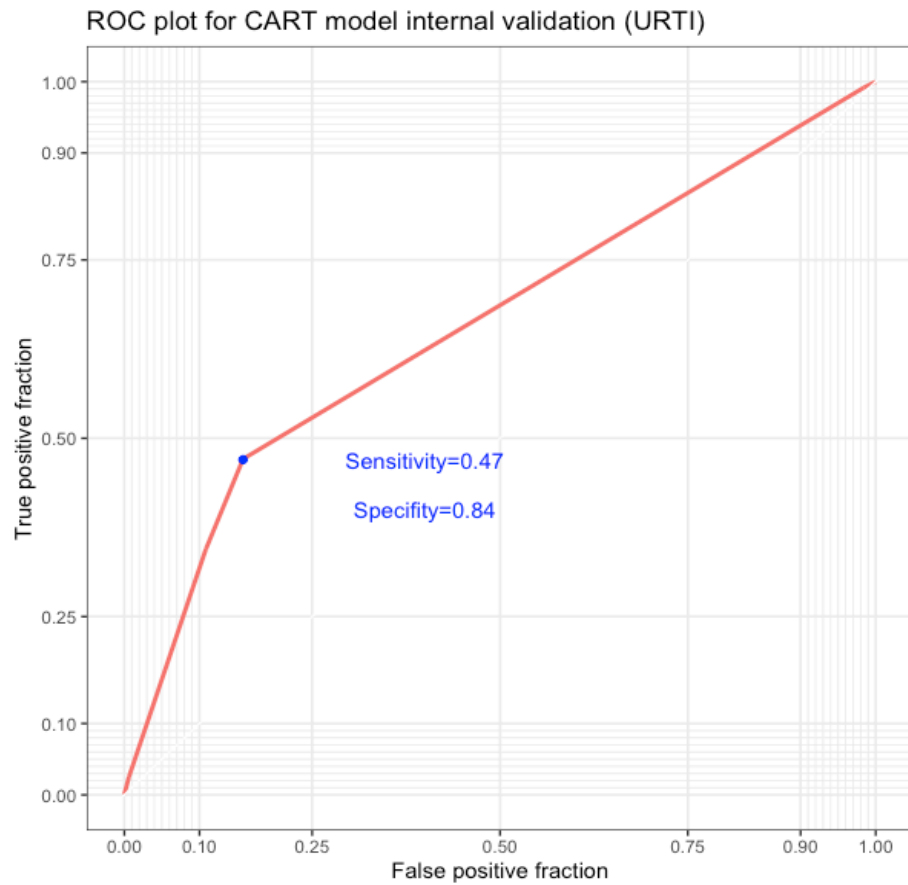


Figure 8.15: ROC curve for internal and temporal validation of CART model (URTI model)

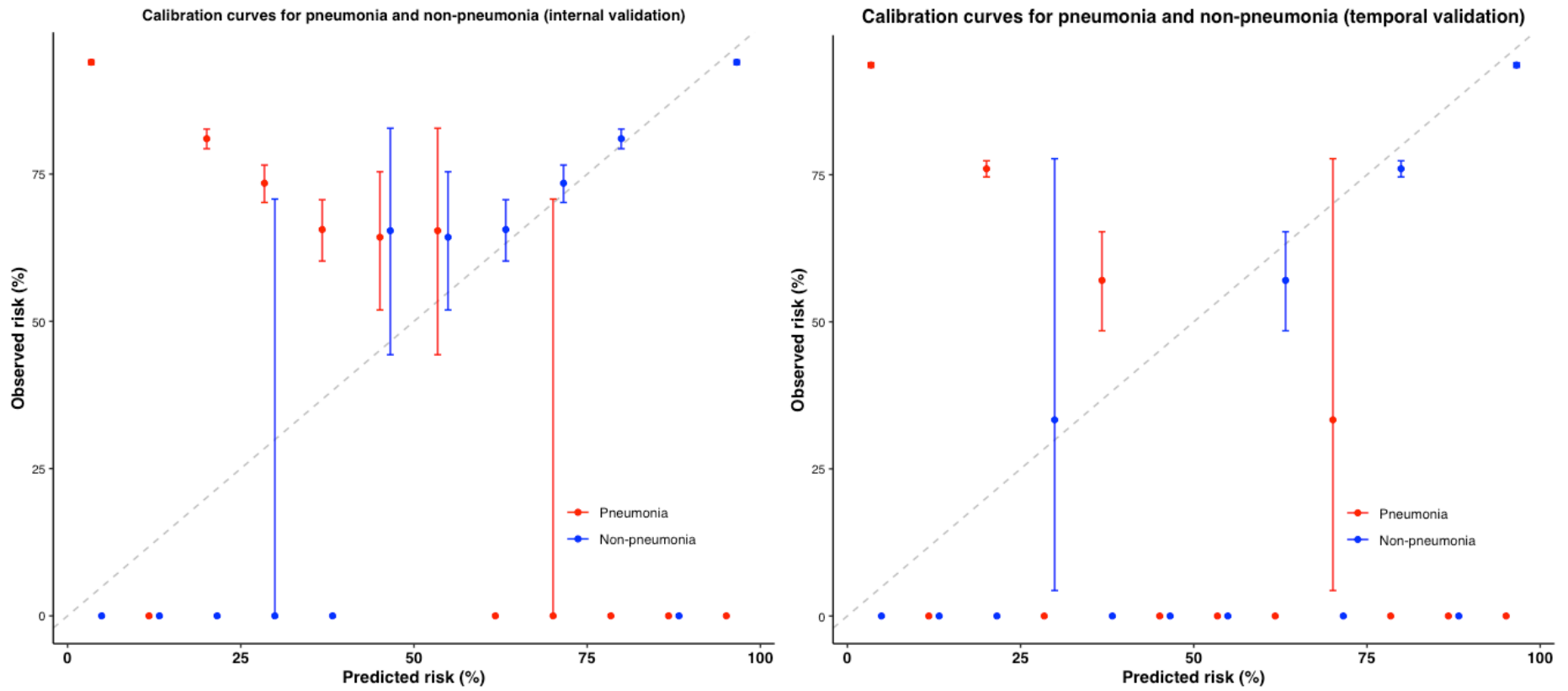


Figure 8.16: Calibration plots for internal and temporal validation of CART (URTI model)

The same four predictors for CART model including, age group, Charlson co-morbidity count, asthma drug use and BMI category were fitted into simple logistic regression model with their effects quantified show in Table 8.16.

Patients reconsulted with pneumonia within 30 days after previous URTI consultations were 0.34 (0.30 to 0.38) times less likely under asthma drug treatment than those free of pneumonia re-consultations. The likelihoods of URTI patients reconsulted with pneumonia within 30 days increased parallelly with age with the highest odds (OR 8.07, 95% CI (7.14 to 9.12)) found among patients older than 85. Similar trend is noted for Charlson comorbidity counts, that is, pneumonia re-consultations were more likely to be found among URTI patients with multi-comorbidities. Pneumonia re-consultations were found 1.85 (1.61 to 2.13) more likely among under weighted URTI patients compared with healthy weight patients.

Both internal and temporal validation performances of simple logistic regression model as shown in Table 8.17, Figure 8.17 and Figure 8.18. Similar to the CART model, low sensitivities (0.02 for internal validation and 0.03 for temporal validation) and high specificities (1.00 for both validations) are noticed.

Discrimination performances with statistical differences compared to the CART model are demonstrated with 0.73 and 0.75 for internal and temporal validations respectively. Simple logistic regression model outperformed CART model in internal and temporal validations. This simple logistic regression model is replicable as shown by the calibration performances measured by H-L test (p value > 0.05 for both internal and temporal validations).

Comparisons of discriminative performances between internal and external validations of developed models through logistic regression and CART were reported in Table 8.18. Apart from logistic regression model for LRTI patients, no significant differences in terms of discriminative performances as measured by AUROC were identified between internal and temporal validations for the rest five prediction models.

Table 8.16: Simple logistic model using same variables from those for CART (URTI model)

	Odds ratio	95% CI	p value
Age group			
Age group (16,35] (Ref)			
Age group (35,45]	2.03	(1.82, 2.26)	<0.01
Age group (45,55]	2.07	(1.85, 2.31)	<0.01
Age group (55,65]	2.92	(2.63, 3.24)	<0.01
Age group (65,75]	3.58	(3.23, 3.97)	<0.01
Age group (75,85]	4.93	(4.43, 5.50)	<0.01
Age group (85,110]	8.07	(7.14, 9.12)	<0.01
Charlson comorbidity count			
Charlson Count (0) (Ref)			
Charlson Count (1)	1.61	(1.51, 1.72)	<0.01
Charlson Count (2)	2.5	(2.30, 2.73)	<0.01
Charlson Count (3)	3.32	(2.97, 3.70)	<0.01
Charlson Count (4)	4.32	(3.71, 5.02)	<0.01
Charlson Count (5)	5.02	(3.97, 6.32)	<0.01
Charlson Count (6)	6.12	(4.41, 8.47)	<0.01
Asthma drug use			
Asthma drug use (Yes)	0.34	(0.30, 0.38)	<0.01
Body weight			
Healthy weight (Ref)			
Under weight	1.85	(1.61, 2.13)	<0.01
Overweight	0.77	(0.72, 0.82)	<0.01
Obese	0.74	(0.68, 0.80)	<0.01
Severe obese	0.72	(0.63, 0.82)	<0.01
Morbid obese	0.90	(0.77, 1.06)	0.21
BMI information not recorded	0.99	(0.91, 1.09)	0.86

Table 8.17: Simple logistic model performance for internal and temporal validations (URTI model)

	Internal Validation	Temporal Validation
Sensitivity	0.02 (0.01, 0.02)	0.03 (0.02, 0.04)
Specificity	1.00 (1.00, 1.00)	1.00 (0.99, 1.00)
Positive predictive value	0.46 (0.33, 0.59)	0.43 (0.34, 0.51)
Negative predictive value	0.91 (0.91, 0.91)	0.90 (0.90, 0.90)
Positive likelihood ratio	8.40 (5.05, 13.99)	6.48 (4.62, 9.10)
Negative likelihood ratio	0.99 (0.98, 0.99)	0.97 (0.97, 0.98)
AUROC	0.73 (0.72, 0.75) ^a	0.75 (0.73,0.76) ^a
H-L test	p-value = 0.75	p-value = 0.49

^a Difference in AUROC between CART and simple logistic regression models is significant (p value<0.01)

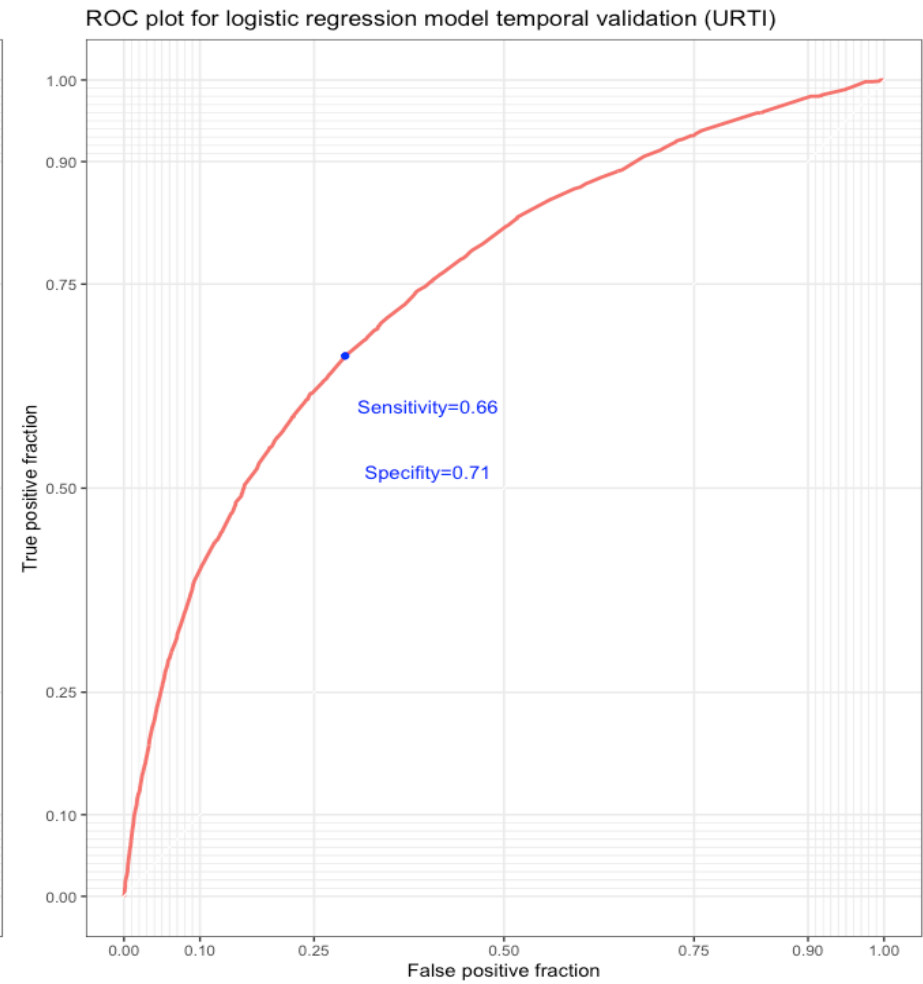
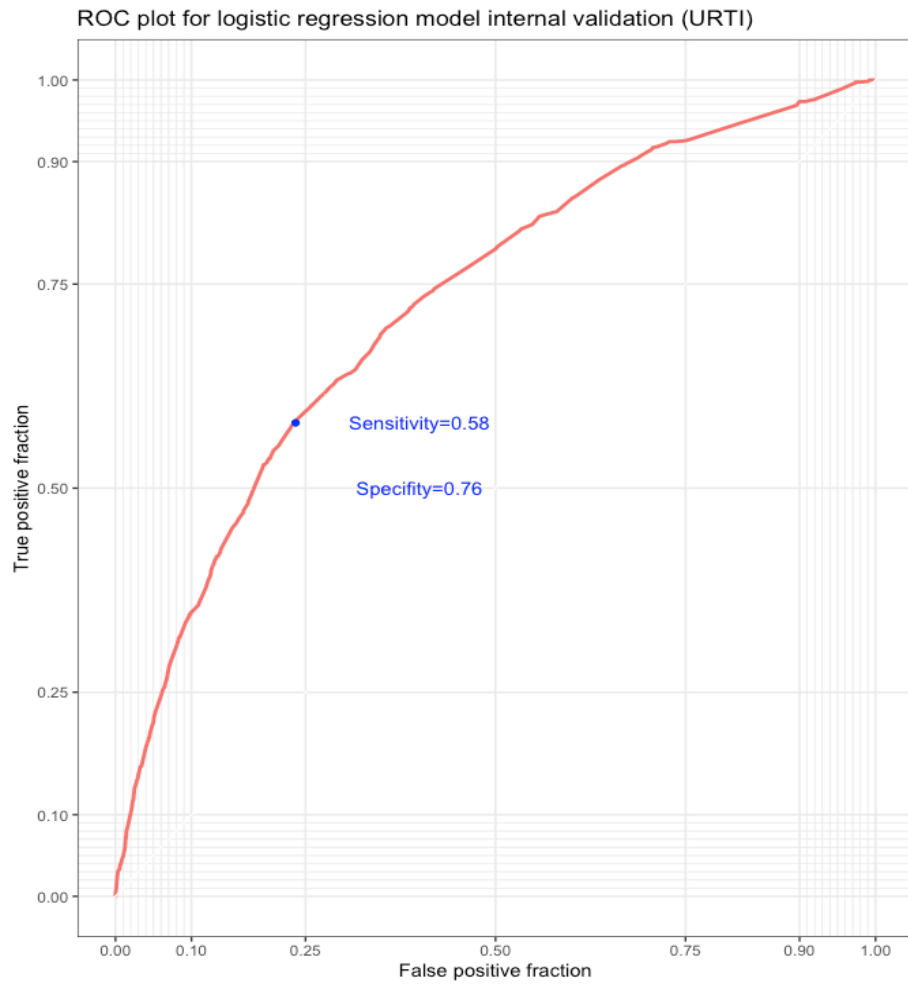


Figure 8.17: ROC curve for internal and temporal validation of simple logistic model (URTI model)

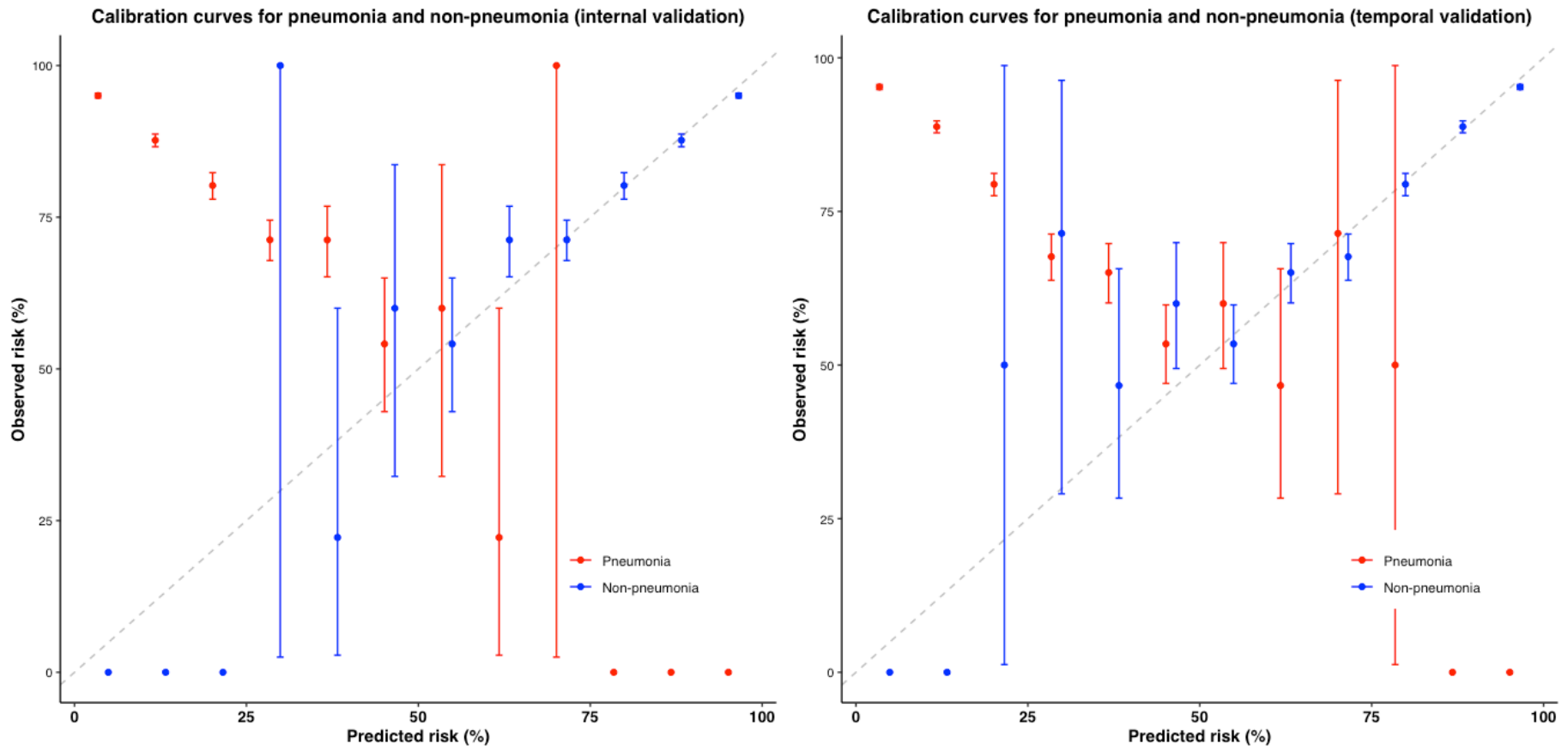


Figure 8.18: Calibration plots for internal and temporal validation of simple logistic regression (URTI model)

Table 8.18: Summary of discriminative performances of developed models as measured by AUROC

Models		Internal validation (AUROC)	Temporal validation (AUROC)
Full model	Logistic regression	0.81 (0.71, 0.74)	0.80 (0.79, 0.81)
	CART	0.70 (0.69, 0.71)	0.68 (0.67, 0.69)
LRTI model	Logistic regression	0.75 (0.73, 0.77)	0.45 (0.41,0.49)
	CART	0.69 (0.67, 0.70)	0.69 (0.67, 0.71)
URTI model	Logistic regression	0.73 (0.72, 0.75)	0.75 (0.73,0.76)
	CART	0.66 (0.65, 0.67)	0.68 (0.67,0.69)

8.1.5 Sensitivity analysis for full model

A subset of data from 2014 to 2017 was deployed for sensitivity analysis. The data set was randomly split into 80% for model development and 20% for the internal with the most recent 25% data used for temporal validation.

The top 15 important variables for three models (random forest, penalised regression models and simple logistic regression model after backward variable selection) are presented in Table 8.19 with results in each process reported in Table A 28, Table A 29, Table A 30, Table A 31 and Figure A 17, Figure A 18, Figure A 19, Figure A 20. The top six important variables identified by three modelling approaches are: age group, chest infection, Charlson comorbidity count, eFrailty index, BMI category and immune system conditions. These six variables were used for medium tree model with five levels of split. The tuning parameter was selected as shown in Figure A 21. In the final CART model, four variables are included: age group, chest infection, antibiotic prescription and immune system conditions.

Table 8.19: Top 15 variables as selected by simple logistic regression, penalized regression and random forest for sensitivity analysis of full model

Logistic regression	Penalized regression	Random forest
Sore throat	eFrailty Index	Age
Rhinosinusitis	Otitis media	Chest infection
Otitis media	Rhinosinusitis	eFrailty Index
Asthma drug use	Sore throat	BMI category
Charlson Count	Cold/ Influenza/ URTI	Season
Immune System Condition	Cough	Charlson Score
Cold/ Influenza/ URTI	Asthma drug use	Age group
Cough	Chest Infection	Smoking status
eFrailty Index	Charlson Count	Charlson Count
BMI category	BMI category	eFrailty category
Diabetes	Immune System Condition	Gender
Clinical test	Multi-Comorbidity	Immune system condition
Age	Age group	Multi-comorbidity
Chest Infection	Diabetes	Clinical test
Gender	Peptic Ulcer	Antibiotic Ever

The CART model is illustrated below in Figure 8.19. Among RTI patients, those who presented with LRTI evidence labelled as ‘chest infection’ were at higher risk of reconsulting with pneumonia in the subsequent 30 days. Within this patient group, those who age above 65 and did not receive antibiotic prescriptions had the highest probability (79%) of re-consultation with pneumonia; patients younger than 66 without antibiotic treatment also had a relatively high probability (61%). For LRTI patients, even if antibiotic prescriptions were issued by GPs, there were two subgroups of patients who were more likely to reconsult with pneumonia in the following 30 days: LRTI patients aged 85 and above and those age between 66 and 85 without immune system conditions.

Both internal and temporal validation performances as shown in Table 8.20, Figure 8.20 and Figure 8.20 demonstrated that this CART model has low sensitivities (0.26 for both internal and temporal validations) but high specificities (0.97 for internal validation and 0.98 for temporal validation). Comparison statistics of development data and temporal validation data is shown Table A 32. This gives model discrimination performances being 0.67 for both internal and temporal validations. This CART model is replicable using recent data as shown by the calibration performance of temporal validation as measured by H-L test (p value > 0.5).

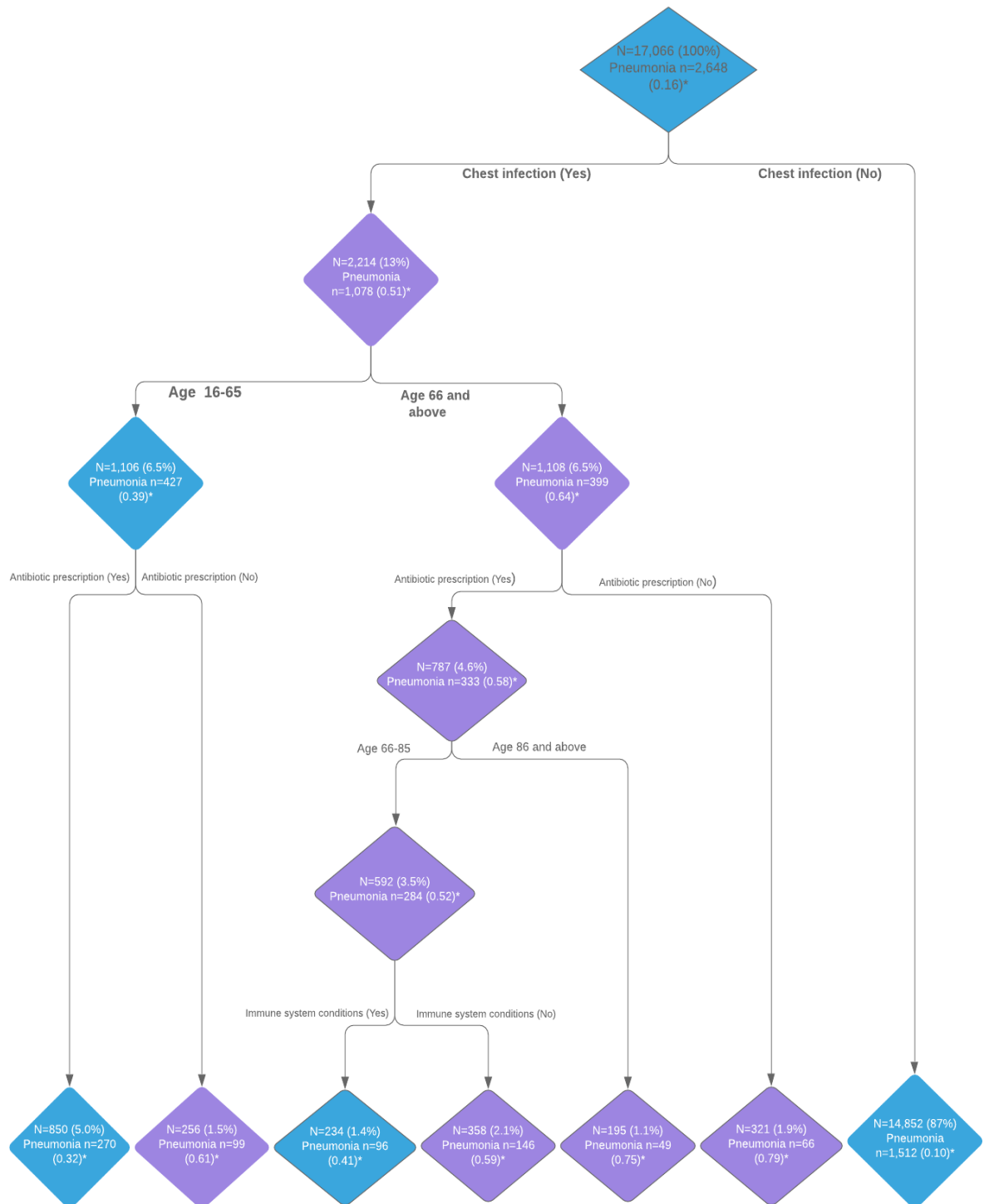


Figure 8.19: CART model sensitivity analysis of full model

* Probabilities of developing pneumonia among patients in individual nodes or leaves

Purple: Majority of patients in individual nodes being pneumonia cases

Blue: Majority of patients in individual nodes being non-pneumonia cases

**Table 8.20: CART model performance for internal and temporal validations
(sensitivity analysis of full model)**

	Internal Validation	Temporal Validation
Sensitivity	0.26 (0.23, 0.30)	0.26 (0.23, 0.30)
Specificity	0.97 (0.97, 0.98)	0.98 (0.98, 0.99)
Positive predictive value	0.65 (0.59, 0.71)	0.73 (0.67, 0.79)
Negative predictive value	0.88 (0.87, 0.89)	0.88 (0.87, 0.89)
Positive likelihood ratio	10.02 (7.90, 12.70)	14.79 (11.12, 19.67)
Negative likelihood ratio	0.76 (0.72, 0.79)	0.75 (0.72, 0.79)
AUROC	0.67 (0.65, 0.69)	0.67 (0.65,0.69)
H-L test	p-value = 1	p-value = 0.88

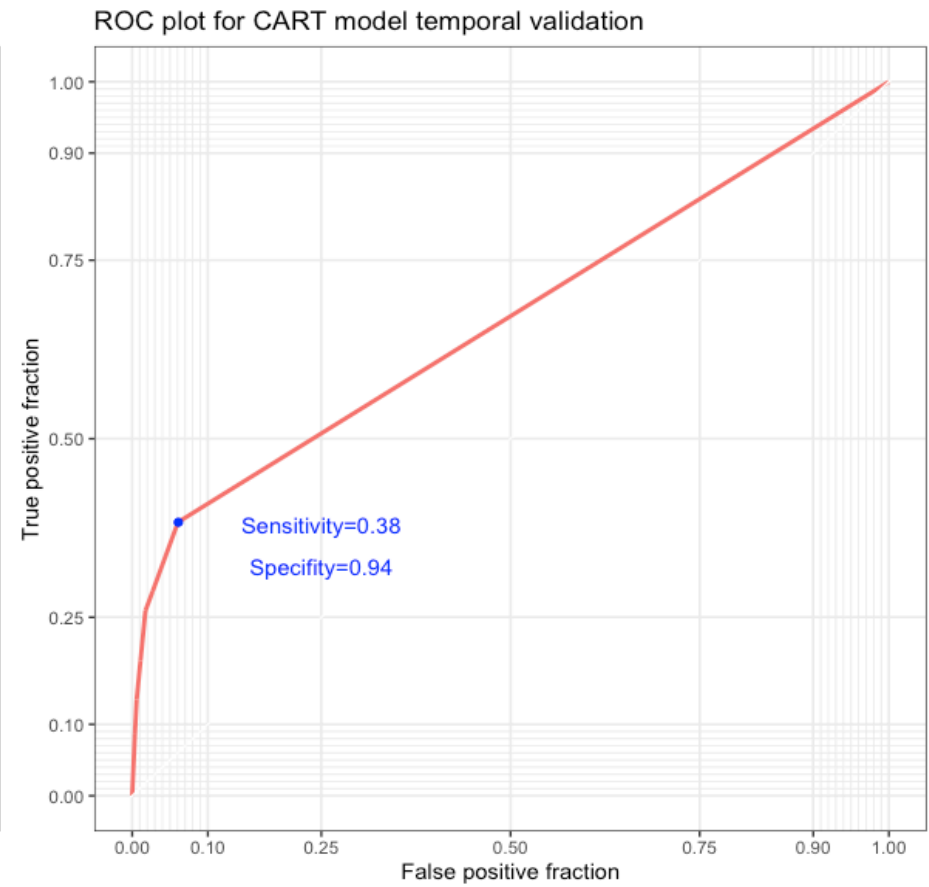
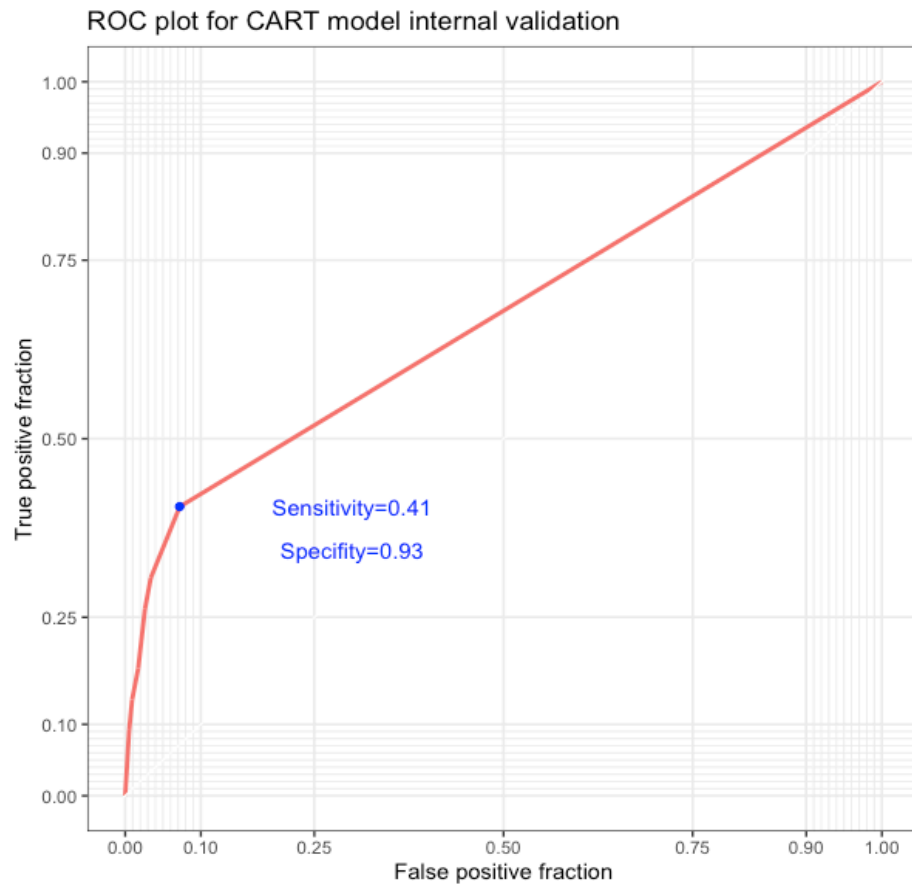


Figure 8.20: ROC curve for internal and temporal validation of CART model (sensitivity analysis of full model)

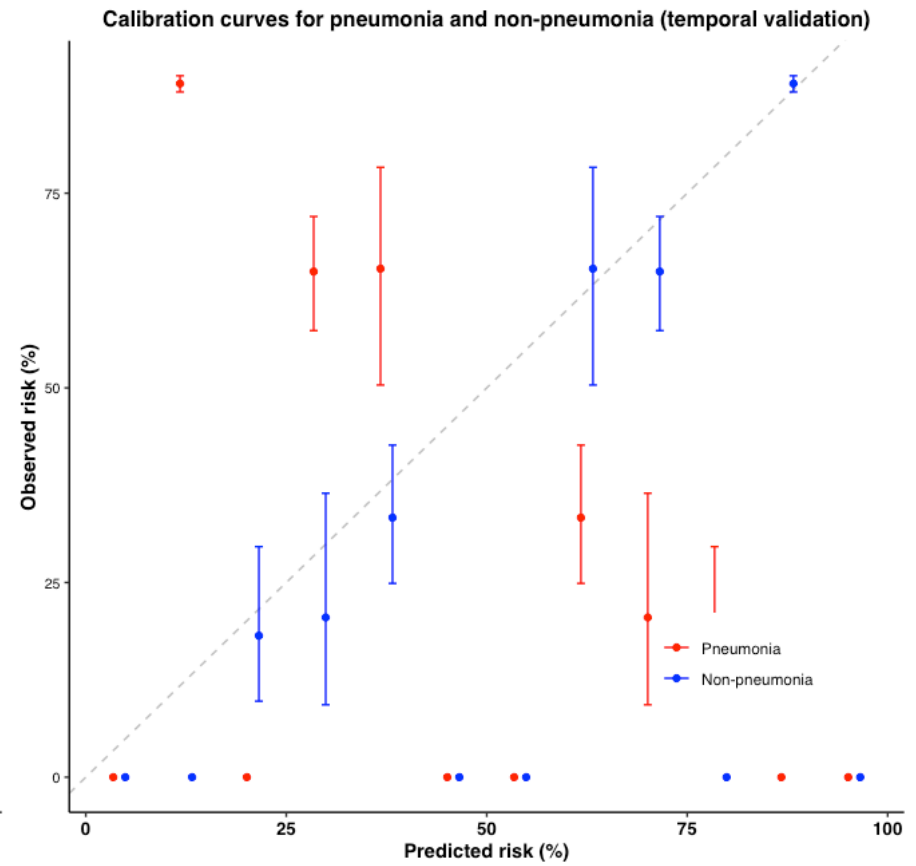
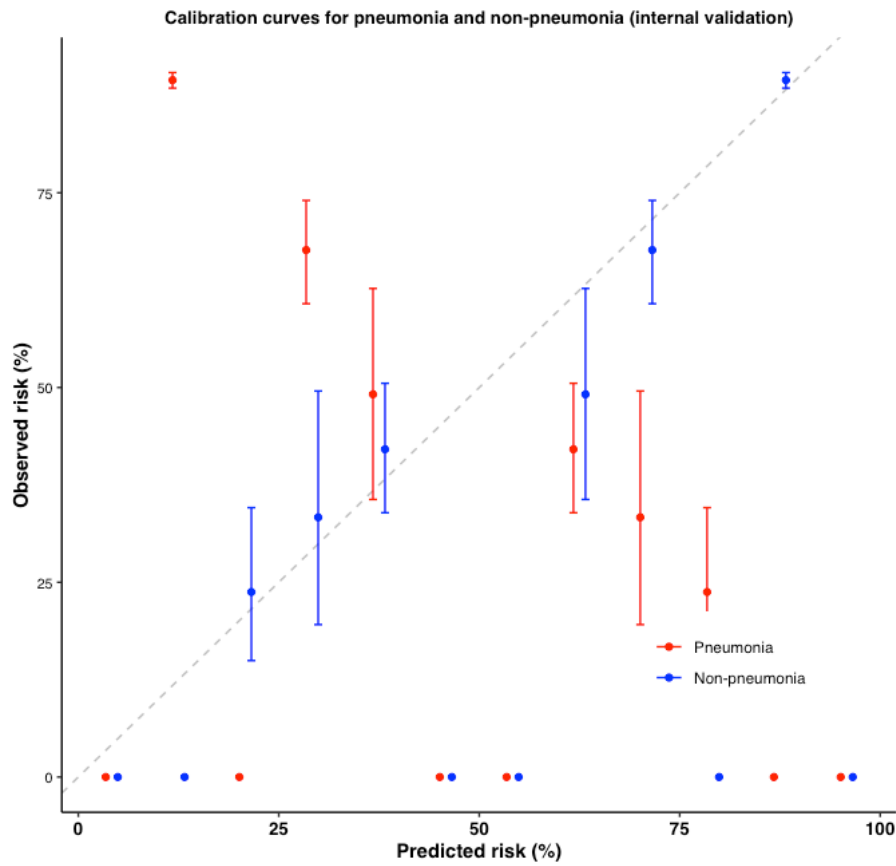


Figure 8.21: Calibration plots for internal and temporal validation of CART model (sensitivity analysis of full model)

The same four predictors for CART model including chest infection, age group, antibiotic prescription and immune system conditions were fitted into simple logistic regression model with their effects quantified show in Table 8.21.

Patients reconsulted with pneumonia within 30 days after previous RTI consultations were 8.05 (7.21 to 8.99) times more likely presented with LRTI symptoms during initial visits than those free of pneumonia re-consultations. The likelihoods of RTI patients reconsulted with pneumonia within 30 days increased parallelly with age with the highest odds (OR 16.22, 95% CI (13.22 to 19.95)) found among patients older than 85. Antibiotic prescriptions have shown to be slightly protective for RTI patients from pneumonia re-consultations (OR 0.91, 95% CI (0.82 to 1.00)). Pneumonia re-consultations were more likely to be found among RTI patients without immune system conditions (OR 0.60, 95% CI (0.53 to 0.68)).

Both internal and temporal validation performances of simple logistic regression model as shown in Table 8.22, Figure 8.22 and Figure 8.23. Similar to the CART model, low sensitivities (0.27 for internal validation and 0.26 for temporal validation) but high specificities (0.98 for internal validation and 0.98 for temporal validation) are noticed. Better discrimination performances with statistical differences compared to the CART model are demonstrated with 0.81 and 0.80 for internal and temporal validations respectively. This simple logistic regression model is not replicable using recent data by the calibration performances measured by H-L test (p value < 0.05 for temporal validation).

Table 8.21: Simple logistic model using same variables from those for CART (sensitivity analysis of full model)

	Odds ratio	95% CI	p value
Age group			
Age group (16,35] (Ref)			
Age group (35,45]	1.75	(1.42, 2.16)	<0.01
Age group (45,55]	2.20	(1.82, 2.67)	<0.01
Age group (55,65]	2.95	(2.45, 3.55)	<0.01
Age group (65,75]	4.79	(4.02, 5.73)	<0.01
Age group (75,85]	7.83	(6.54, 9.42)	<0.01
Age group (85,110]	16.22	(13.22, 19.95)	<0.01
Immune system condition (Yes)	0.60	(0.53, 0.68)	<0.01
Antibiotic prescription (Yes)	0.91	(0.82, 1.00)	0.05
Chest infection (Yes)	8.05	(7.21, 8.99)	<0.01

Table 8.22: Simple logistic model performance for internal and temporal validations (sensitivity analysis of full model)

	Internal Validation	Temporal Validation
Sensitivity	0.27 (0.23, 0.30)	0.26 (0.23, 0.30)
Specificity	0.98 (0.97, 0.98)	0.97 (0.97, 0.98)
Positive predictive value	0.69 (0.63, 0.75)	0.63 (0.57, 0.69)
Negative predictive value	0.88 (0.87, 0.89)	0.88 (0.86, 0.89)
Positive likelihood ratio	12.35 (9.95, 15.91)	9.20 (7.26, 11.66)
Negative likelihood ratio	0.75 (0.71, 0.78)	0.76 (0.72, 0.80)
AUROC	0.81 (0.79, 0.83) ^a	0.80 (0.79,0.82) ^a
H-L test	p-value = 0.57	p-value = 0.01

^a Difference in AUROC between CART and simple logistic regression models is significant (p value<0.01)

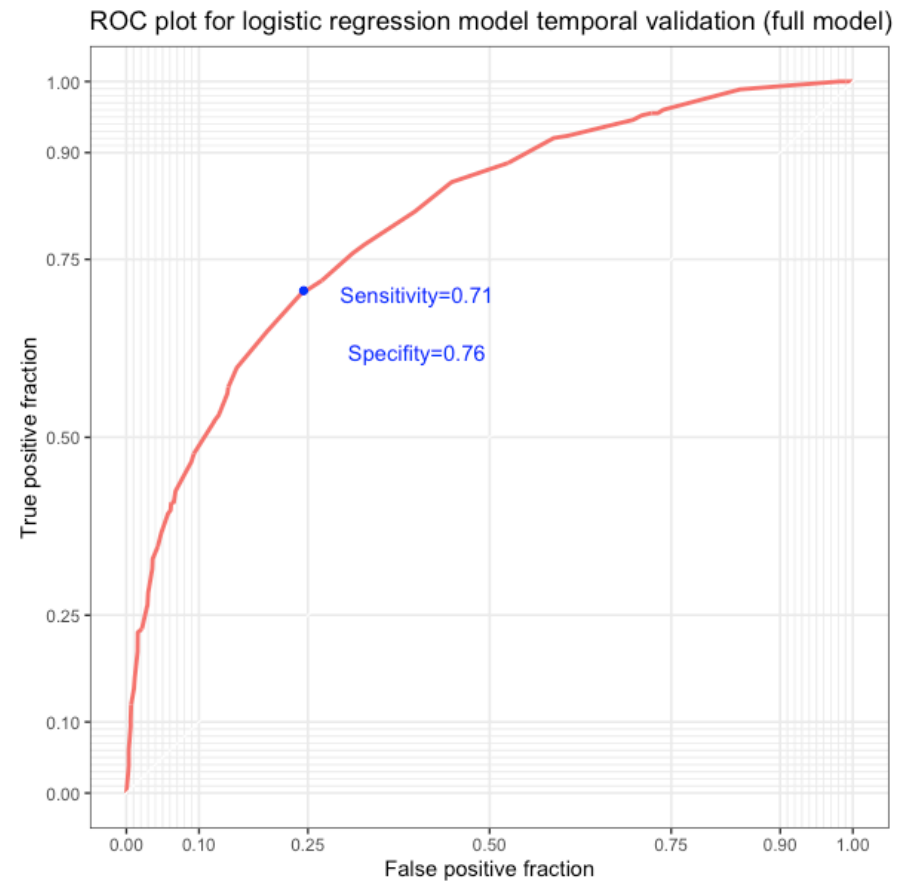
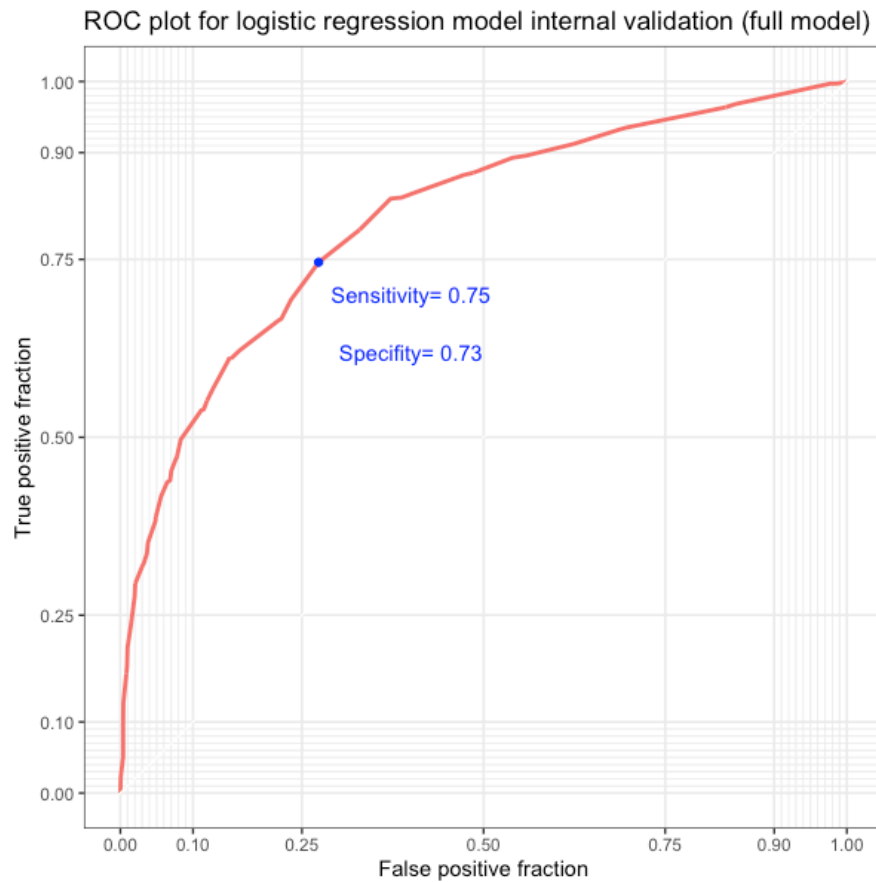


Figure 8.22: ROC curve for internal and temporal validation of simple logistic regression model (sensitivity analysis of full model)

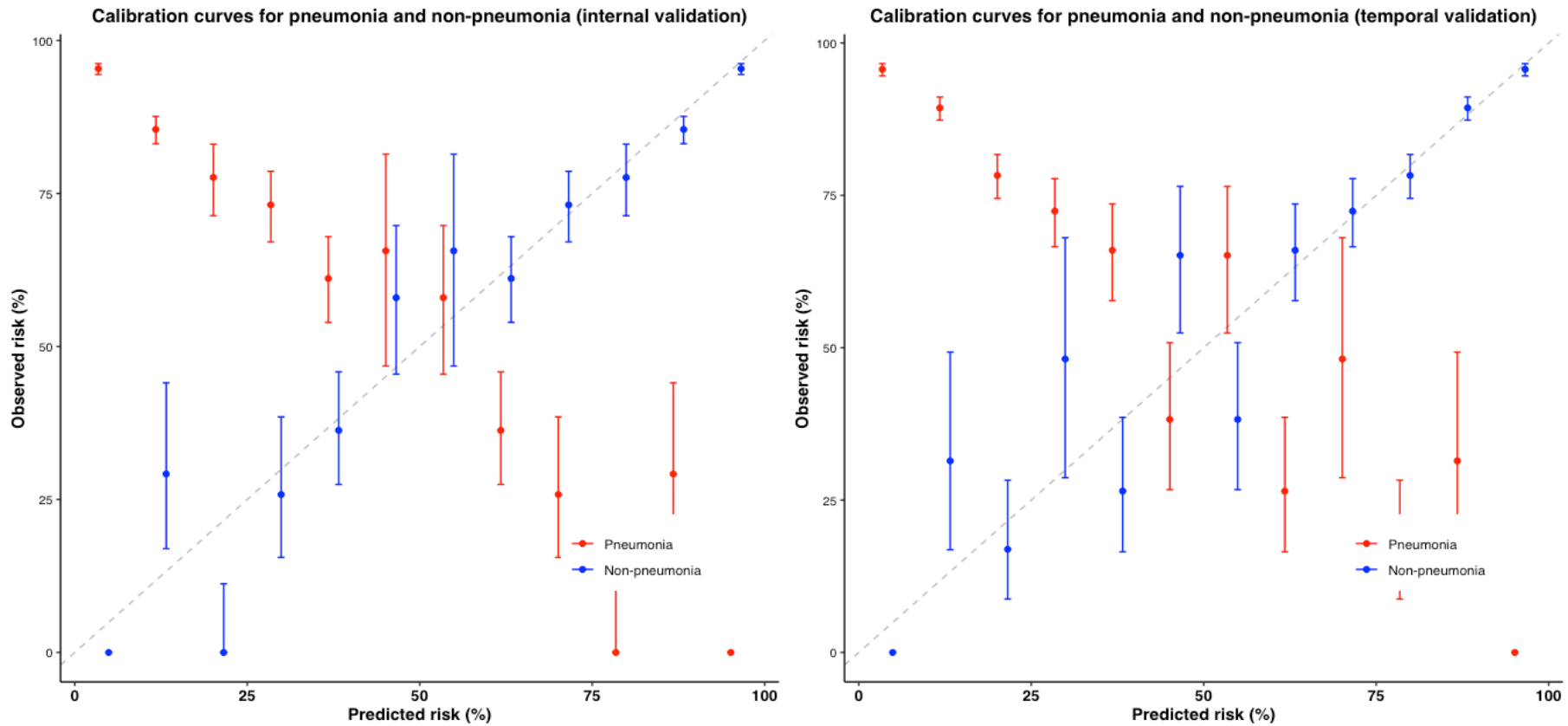


Figure 8.23: Calibration plots for internal and temporal validation of simple logistic regression model (sensitivity analysis of full model)

8.2 Discussion

8.2.1 Main findings

8.2.1.1 Patient's profile for RTI patients who reconsulted with pneumonia

RTI patients who presented with LRTIs generally had higher odds of reconsulting with pneumonia with several characteristics and antibiotic prescription shown to be protective for pneumonia re-consultations.

For LRTI patients aged over 85, even if antibiotic prescriptions were issued during clinical consultations, the management outcome was not always optimal. This may not only indicate that current antibiotic empirical treatment plan does not work well at advanced ages, but this might also suggest that the LRTI presentations could be the early onset of other underlying health conditions. Antibiotics may often be prescribed during end-of-life care. Antibiotic management failure was also found among LRTI patients with multiple comorbidities. Possible reasons could be that complicated or atypical RTIs were presented among those patients or deteriorated existing conditions contributed to the onset of pneumonia later on. Therefore, care bundles which investigate the original causal factors for LRTIs or tailor to intensified management of the long-term conditions meanwhile treating LRTIs could be considered.

Study results from LRTI patients aged 16 to 65 group suggested that the onset of LRTI among patients without apparent risk factors could be viewed as an early warning message that the patient's general health needs to be checked further. This is because pneumonia disproportionately affects the very old and young, this age group should not be the conventional risk group. Plus, these patients were counterintuitively identified as risk cohort without presenting common risk factors of pneumonia.

The model for URTI patients did not perform entirely well, nevertheless, it still suggests that being underweight could be the sign of deteriorated health conditions

(Flegal et al., 2005), especially for patients living with chronic diseases. As shown in the leaf nodes in the CART model for URTI patient, the ones with the lowest re-consultation risk were found among patients who were relatively young (16 to 75 years old) with more than one comorbidity and under asthma treatment. This together with other apparent counterintuitive findings indicate that defensive approaches probably had been adopted during clinical practice in the UK primary care system—patients presented with common risk factors were treated more conservatively whereas apparent healthy patients may need prompt treatment or further investigations. Future investigative studies designed based on major risk factors identified from this study could be conducted with high efficiency. For example, C-reactive protein test could be offered to adult LRTI patients who are younger than 65 without apparent conventional risk factors of developing pneumonia to assist clinical decision of antibiotic treatment. If test results do not support common bacterial pulmonary infections, other clinical investigations then should be issued to find out the underlying reasons for the onset of LRTIs i.e., lung cancer. This could also contribute to healthcare policy development such as fast referral care pathways in respiratory medicine. Additionally, the relatively low numbers of pneumonia cases in each leaf node of CART model for URTI patients in comparison with descriptive statistics that more than half of pneumonia patients had previous URTI consultations, suggest that the extensive heterogeneity of underlying patterns responsible for URTI patients' pneumonia re-consultations in the community. This together with constant high specificities reported across all models through model performance evaluations for URTI patients, indicated that disease management for URTI in the UK primary care was generally at satisfactory level.

The sensitivity analysis shows much the same results compared with planned analysis using 16-year data. But the calibration accuracy for CART model supports the replicability of model using up to date information even at the expense of sample size.

In summary, the prediction modelling results suggested that age, comorbidity and presentation of LRTIs are the main risk factors for RTI patients regarding short term pneumonia re-consultations in the UK primary care settings. These risk factors are

similar to the ones for incidence CAP, which broadly reflected the weakened lung immunity (Almirall et al., 2017, Chalmers et al., 2017). The dominant risk factor identified from this study was LRTIs suggesting acute inflammatory responses were triggered (Mizgerd, 2008). And if such irritant was caused by bacterial pathogens, this means that decreased mucociliary clearance efficiency of upper respiratory tract allows overloaded virulence reach the lung or less efficient immune system fails to prevent the onset of infections (Brown, 2012, Mizgerd, 2008). Another significant risk factor was age. The reasons why increasing age has an influential impact on adult incident CAP are multifactorial: aging in itself indicates the accumulation of certain hazardous effects of coexisting conditions; it also has a direct impact on decreased innate and adaptive immunity (Baxendale and Brown, 2012, Sahuquillo-Arce et al., 2016). Such deteriorated general health was reflected by the predictive models as measured by Charlson comorbidity count in this study. This probably suggests that diagnosed long term health conditions may represent the increased risk better among RTI patients for the onset of CAP better compared to other measurements such as frailty. When it comes to other infectious conditions, frailty was reported to be associated with higher risk of sepsis (Gulliford et al., 2020). However, certain well-known risk factors for CAP were not identified as strong predictors such as smoking (Baskaran et al., 2019). This together with other counterintuitive results such as asthma drug use and immune system conditions became protective factors against pneumonia re-consultations may results from conservative management strategies were adopted during initial RTI treatment. Therefore, such results indicate that more attention should be paid to apparent healthy patients during clinical practice especially when LRTIs are presented.

8.2.1.2 Variable selection

Prediction modelling could be considered to identify a model, by and large, to answer predefined research question and the role of information as expressed by individual variables or variables serves to reduce uncertainty. These entail the rationale of variable selection to filter out several variables with most predictive values to capture the majority aspects of research question of interest. In contrast to other de-dimensional techniques like principle component analysis (PCA), machine

learning algorithms deployed for variable selection take the target into consideration without generating new variables which are linear combinations of old variables (Peacock and Peacock, 2011). That is, PCA is valid for variable selection only if the most predictive variables are those have the most variations in them, which is not always the case.

Given the different results presented in this study during the variable selection process, it is crucial to understand how these algorithms behave in order to critically appraise the output and, to make sense out of the data. First, predictive information is more likely to be identified by various machine learning algorithms or similar algorithms with slightly different calculation criteria. During variable selection for the full model, LRTIs denoted by ‘chest infection’ was selected as an important variable. Also, as illustrated in chapter seven, sex is chosen over the travel class for the initial split for classification tree approach to predict survival status of passengers on Titanic by various splitting criteria. Similar results were noticed during variable selection for ‘chest infection’ if predictive accuracy was chosen to rank variable importance instead of mean decreased in Gini index. There were two variable ranking techniques in random forest: mean accuracy reduction and Gini impurity reduction (Chicco and Rovelli, 2019). Variable importance measured by mean accuracy reduction is based on the comparison of accuracy decreased by removing one variable to that obtained using all variables- the greater the accuracy decreased, the more important the variable is. Similar for mean decreased Gini impurity, the most important variable contributes to the largest Gini impurity reduction (Chicco and Jurman, 2020). In this study, mean decreased Gini was chosen as the prediction models aim to find more homogenous subgroups of RTI patients who reconsulted with pneumonia and those did not.

During variable selection, even if certain data transformation procedures to centre or scale the variables were not performed, cautions were given to inspect skewed data and the ones with high proportions of unique values so called as ‘near zero-variance variables’ (Kuhn, 2008). In this study, three types of algorithms were adopted to perform variable selection. This was aimed to guide variable selection when similar information was presented with various forms, for example when ill health was

measured using the eFI, frailty categories, individual long-term medical conditions and multi-comorbidities. It is expected that same variable was selected but with various orders in terms of variable importance. Generally, simple logistic regression and penalized regression favour categorical variables over their continuous forms as they quantify variable importance by the absolute values of the coefficients that is, the effect size of variables contribute to the outcome interests. If the same information such as age was included as continuous variable, then the effect size is 'attenuated' by presented as average effects on the outcome of per unit change compared to include the same information presented as age group. This also explained why tree-based models favour continuous variables or categorical variables with more categories because they are able to offer more options to split the tree into more homogenous subgroups as continuous variables could be considered as a special kind of categorical variable with infinite number of possible categories.

Random forest was adopted adding information to penalized regression results. Two hyperparameters were used to optimize the algorithm: Ntree and Mtry. Ntree serves to identify the possible number of decision trees using different random selection (with replacement) of Mtry variables which aims to reduce the complexity of prediction model in avoidance of overfitting. The number of trees was identified by OOB misclassification errors to facilitate variable selection at acceptable level of error meanwhile at the ease of computation efficiency. The voting procedure share much the same concept of boot strapping and stepwise selections which could be trace back to the law of probability. The classic example to illustrate Gaussian distribution is to toss the coin multiple times, when the repetition reach to sufficient times, a normal distribution with confidence interval to quantify the most likely range of true value is generated. Here, we 'plant' the decision trees with enough times, then the majority of voting would be presented as the model lies in the 'confidence interval'. When the OOB errors arrived at a stable and acceptable level, then the minimum number of trees could be considered as enough number for modelling in terms of the size of the forest.

8.2.1.3 Model specification

Model specification of prediction model in this study adopt CART and simple logistic regression. CART models are able to capture the interactions between predictors which are less straightforward as expressed by interaction terms for simple logistic models (Karaca-Mandic et al., 2012). Meanwhile, the overall predictive effects of included variables could be quantified by simple logistic regression model. Even if CART models do not produce ‘effect sizes’ compared to simple logistic regression models, the validity of CART models still could be partially explained by the estimation of simple logistic regression models. For example, the estimations of odds ratio of ‘chest infection’ supports the first level of split on this variable for full model as well as sensitivity analysis results. Also, the overlapping confidence intervals between 36-45 and 46-55, 46-55 and 56-65 age groups as shown by the sensitivity analysis indirectly support that patients who were younger than 65 was classified on the same node during next level of splitting. Given that the prevalence of pneumonia went beyond 10% due to the sampling strategy of this study which overestimate the relative risk (Zhang and Kai, 1998), nevertheless, the direction as well as the nature of confidence interval should not be changed if further adjustment is made.

8.2.1.4 Model performance

Prediction model in healthcare could be viewed as non-invasive medical device, for which the concept of measurement system analysis (MSA) (Montgomery, 2020) could be applied to evaluate its performance. That is, there are generally four aspects to consider: precision, accuracy, stability and bias. The precision of the estimation largely depends on the size of meaningful data used for model specification usually reflected by the confidence intervals for certain measurement outputs. The accuracy of prediction model was assessed by two common measurements: if the model could accurately identify cases out of real cases and non-cases out of real non-cases- the overall discriminative performance is quantified by AUROC; if predicted cases accurately align with the real cases and predicted non-cases with real non-cases- calibration test. The stability of developed model is quantified by model validation: if

the model performs stably within the cohort whose information was adopted to develop the model-internal validation; if the model still performs well over time-temporal validation; if the model could be applied to same cohort somewhere else-spatial validation or generalized to another similar but completely different cohort-external validation. Bias normally refers to the systematic error, by quantifying which could estimate if the bias is sufficient large to distort the conclusions or adjust measurement required during real life application. Even if quantifying the bias is not always feasible, by identifying the source of bias still contribute to restrict the scope of application or the overarching conditions under which the model is valid.

For prediction modelling, model performance tests generally emphasis on accuracy and stability. In this study, simple logistic regression models outperformed CART models with same variables as assessed by AUROC. However, this does not necessarily mean that CART models are worse than simple logistic models, rather the usefulness of a prediction models heavily depends on how efficient to answer the research question at hand.

It is noticed that the shapes of ROC curves for CART model differ from those generated from simple logistic regression models. The possible explanation is that simple logistic regression models assign individual predicted risk estimation to each patient, whereas for CART models, patients classified into the same leaf node were allocated with the risk estimation of that group. This might also explain why logistic models out-performed CART models in terms of discriminative performance for the same groups of patients during internal validations. Because logistic regression is able to assign individualized risk probabilities to each patient in the validation cohort which derives from the same parent cohort of training dataset. In other words, prediction models quantified by logistic regression are generally expected to fit better for test dataset during internal validations comparing to those from CART. However, logistic regression models may not continue to have better discriminative performances than CART models as indicated by temporal validation for LRTI model. CART model is developed to identify homogeneous groups which allows to capture the key features of the problem at hand as reflected by the first several splitting criteria. As long as there were no significant changes among identified key

features, major discrepancies are not expected between internal and temporal validations for CART models. On the other hand, logistic regression models are more sensitive to any changes of variables included in the model. As result, the secular trends of CAP as reported in chapter five might partially be responsible for such less stable performance of logistic regression model for LRTI patients over time. Such different behaviours between logistic regression and CART were further reflected by the ROC curves: the ones for simple logistic regressions are smoother than those from CART which are dominated by homogeneous groups. Similar results were observed when comparing 15 machine learning approaches using the same data set and cross validation criteria in Table A 16: not only rough ROC curves for tree-based models were displayed in comparison with simple logistic regression model, but less smooth ROC curves from CART models with fewer levels of split (coarse and medium tree) were also noted in contrast to the one from fine tree. This may suggest that CART model could contribute to identify the homogenous groups that dominate model performance which is difficult to be captured by simple logistic regression models. Future prediction modelling research could adopt triangulation approaches layered one on another to take advantage of complementary advantages of various methods especially for real world data that were not purposely collected for specific studies.

In this study, full model of CART together with the one from sensitivity analysis have shown that patients without LRTIs were at the lowest risk group of pneumonia re-consultation, therefore high specificities were identified from model performance evaluation. Plus, the CART model for URTI patients also supports such high specificity results, that is, most patients free of LRTIs were classified into low-risk groups in terms of pneumonia re-consultation.

Discriminative performance as displayed by the ROC curve essentially illustrates the trade-off between sensitivity and specificity and could be used to locate the optimal cut-off value based on clinical context (Florkowski, 2008, Ekelund, 2012).

Therefore, how well the model can deliver the same function as individual case reports? This question is answered by model performance parameters depending on the purpose of developed model in terms of clinical usefulness. If the model is

mainly deployed for screening purposes, apart from discrimination performance as quantified by AUROC, specificity and negative predictive value are the key indicators to see if the model function well given predefined acceptable range relating the specific clinical scenarios i.e., 80% to 90%.

8.2.2 Strength and limitations

To our knowledge, this is the first study applying both machine learning and conventional modelling approaches to routine primary care EHRs to investigate pneumonia patient profile after initial RTI consultations in a large general population. The study results have proven the feasibility of deploying machine learning approaches to explore the patten from ‘real world’ data.

Data driven and machine learning modelling approaches require less expert input during variable selection comparing to conventional statistical modelling techniques. Also, prediction models generated through machine learning algorithms have the potentials to identify novel disease patterns that could be easily missed out by expert experience. Plus, machine learning techniques allow inclusive information on patient’s demographic characteristics, lifestyle, environmental exposures, health conditions, medication prescriptions, disease preventive interventions, clinical management and clinical investigations to be investigated, which makes the use of available data.

In this prediction modelling study, the use of both machine learning algorithms and conventional modelling methods followed the process of so called as ‘triangulation’ (Heale and Forbes, 2013). The analytical approaches were consistently applied to the whole sample as well as subgroups of sample. Systematically examining research question with different statistical procedures allowed critical appraisal of integrated results and provided a more comprehensive understanding of research topic.

Given that the various combinations of patient characteristics are indicating different underlying reasons for pneumonia re-consultation among RTI patients, clinical utility analysis like decision curve analysis (DCA) (Vickers et al., 2016) was not conducted.

This is because the data does not contain detailed information on the exact date if patients went through specific clinical investigations nor the corresponding results. Also, prompt treatment like antibiotic treatment sometimes is sufficient to prevent undesirable outcome, whereas for other group of patients, anti-infectious treatment may not enough to resolve the clinical symptom unless primary underlying disease is managed.

There are several study designs proposed to deal with confounding variable effects in research using observational data including quasi-experimental design (Campbell and Cook, 1979, Posternak and Miller, 2001) and propensity score methods (Austin, 2011). By using these study designs in observational data research, the effects of covariates could be investigated and adjusted during analysis which shares the same intension of randomized control trials with the baseline covariates effects elucidated by study design and randomization. Due to the restrictions on the accessibility of whole CPRD data set of our licence, such study designs were not feasible. Nevertheless, stratified sampling strategy based on general practice and study years has minimized the possible variations owing to data recoding behaviour at general practice level over time, which is less relevant to daily clinical practice at individual level.

During temporal validations, part of the data used for validation was adopted during model development. Such ‘information leakage’ is not optimal for model validation as it does not keep training and test data entirely separate (Steele et al., 2018). However, the study results have already demonstrated that it is able to generate meaningful outcome as long as updated or new information is introduced during validation process. Some models are shown to be lack of replicability, which indicates there were significant chronological trends in terms of disease management, healthcare data documentation, code drifting have happened in the UK primary care system even within four-year period such changes in antibiotic prescriptions addressed in chapter four.

8.3 Conclusion and future implications

The study shows the potentials of using machine learning to implement such techniques to primary care EHRs in the context of pneumonia re-consultation within RTI patient cohort. Compared to conventional modelling approach, machine learning algorithms are able to produce an array of novel and holistic models with wider range of individual health care information. Machine learning shows the value of being able to disentangle the complexity of overlapping information carried by included predictors which could serve for tailored disease management, conducting exploratory analysis, identifying underlying disease pattern, generate research hypothesis even clinical trial recruitment (Weng et al., 2017, Grosios et al., 2010).

In order to translate prediction model into clinical utility, the validity of machine learning algorithms should be established through external validation and replication. This would require transparent reporting of study population characteristics, how healthcare information was constructed into predictors, essential elements of variable selection process as well as model development exercise.

Chapter Nine : Overall discussion and conclusions

This chapter begins with a summary of the main findings of this research, followed by a general discussion and reflection on the thesis, presenting further conclusions and discussing the implications for future research.

9.1 Summary of main findings

This thesis comprised four inter-related studies to offer a coherent understanding of prediction modelling in the context of CAP following RTI consultations in primary care settings. The initial epidemiological study analysed primary care electronic health records data from 102 general practices in England between 2014 to 2017. An annual relative reduction rate (RRR) for total antibiotic prescription of 6.9% was found from 2014 till 2017. The decline in antibiotic prescriptions for broad-spectrum β -lactam showed a faster rate of decline with RRR being 9.3% per year. Overall, more than half (54.1%) of antibiotic prescriptions were not recorded with specific clinical conditions or informative diagnostic codes. Respiratory conditions remained the most frequent indications for antibiotic prescriptions associated with informative medical codes and showed the greatest reduction in prescription rates.

The following epidemiological study aimed to investigate the secular trends in the incidence of pneumonia and related respiratory conditions using data from the UK general practices contributing to CPRD from 2002 to 2017. Four respiratory conditions were evaluated: clinically-diagnosed pneumonia, clinically-suspected pneumonia defined as antibiotic treated chest infection, influenza pneumonia and pleural effusion. Annual percentage change (APC) estimated from joinpoint regression model was adopted to quantify the chronological changes and identify significant turning time points over the 16 years of follow up. Clinically-diagnosed pneumonia incidence was found to increase over time with an accelerated trend with an APC of 5.1% after 2010 compared to that of 0.3% before 2010. For clinically-suspected pneumonia, an overall contemporaneous trend with average increasing rate being 3.8% from 2002 to 2008 whereas a faster decreasing rate of 4.9% thereafter until 2017. Influenza pneumonia increased in the epidemic year of 2009. Pleural

infection rates remained to be static over the study years. The study results suggested that antibiotic prescribing practice together with clinically coding behaviour partly accounted for the apparent increase in clinically diagnosed pneumonia in primary care settings.

In the systematic literature review of current research evidence of prognostic factors of community-acquired pneumonia (CAP) published between 1st January 2010 and 20th January 2020, 30 individual prognostic studies and 16 systematic reviews were included. Individual prognostic studies were assessed for risk of bias. Only 12 studies were rated as being at low risk of bias. Overall, 33 prognostic factors for CAP were identified which could be broadly categorized into six groups: patient general impression, clinical test biomarkers, medication utilization, lifestyle and environmental exposure, preventive procedures and co-morbid conditions. Conflicting results concerning the predictive values of C-reactive protein (CRP), weight gain/ obesity, flu and pneumococcal vaccinations for CAP were reported. The main sources of bias were attributed to statistical methodology, inadequate reporting, lack of control of confounding variables and lack of representative study populations. Understanding the contextual information on how prognostic studies were implemented is essential to explain the prognostic value of identified predictors and provide possible candidate variables that could be potentially included in prediction modelling study.

Based on these study results together with the enriched primary care information documented in CPRD, an inclusive approach was adopted for candidate predictors. Machine learning algorithms together with conventional modelling methodology were employed for variable selection to identify potentially important predictors for the final models. In general, patients presented with LRTIs were at the high risk of reconsulting with pneumonia in the following 30 days. Antibiotic prescription has shown to be protective for RTI patients in terms of near future pneumonia consultations. Among patients presenting with LRTIs, those aged over 85 years remained at higher risk of reconsulting with pneumonia even if antibiotic prescriptions were offered during the consultation; those aged between 76 to 85 with two or more comorbidities as measured by Charlson comorbidity index were also

identified as a risk group despite antibiotic prescription. For patients presenting with LRTIs, but where clinical discretion did not lead to antibiotic management, those who were younger (16 to 65 years old) without asthma nor immunosuppressant drugs appear to have higher risks of pneumonia re-consultation within 30 days; whereas for patients older than 65 the risk would persist independent from other risk factors. However, we caution that allocation to antibiotic treatment as well as other disease management procedures was not randomized and confounding by indication might account for counter-intuitive findings. Meanwhile, for patients presented with URTIs only, the risk of re-consulting with pneumonia within 30 days was found among patients younger than 76 years who were not treated with asthma drugs but diagnosed with more than three long term conditions being under weight. The study results suggest that current management plan of URTIs in the UK primary care was generally appropriate. Prediction modelling results of LRTI patient indicate that more attention should be paid to subgroups of LRTI patients to find out underlying reasons responsible for the onset of LRTIs. Based on study findings, machine learning techniques are able to identify novel disease pattern compared to conventional approaches, which could be used to generate research hypothesis, individualized research design for inventory clinical trials.

9.2 General discussion and reflections on the thesis

Artificial intelligence (AI) or machine intelligence in healthcare, has potential to transform the landscape of medicine and medical research (Matheny et al., 2020b). The major difference between human intelligence and artificial intelligence is that the former is a type of individual intelligence whereas the latter is network based intelligence which may sometimes perhaps be observed in the natural world, as among ant colonies rather than human beings (Sibisakkaravarthi and Subramaniam, 2017). Therefore, direct comparisons between human intelligence and artificial intelligence are generally unnecessary. The essential attributes for this type of intelligence are big data, modelling algorithms and computational technology- the historical trajectory of which is denoted by Moore's law (Moore, 1965). Alternatively, the same concept could be captured as information, instruction commands and energy. That is, the realization of artificial intelligence is a process of

deploying energy to process information under the guidance of algorithms. The discussion of the inter-related research studies of the thesis is integrated into the implications of these three aspects to research.

Energy and Moore's law

Moore's law is not a physical law but a classic example of a prediction model, which is adopted to guide work planning and set targets for professionals in the semiconductor industry. The model states that the number of transistors on a microchip doubles every two years and this growth rate was projected to last for decades (Moore, 1965). From the most primary form of computer-abacus to the current high-performance computer, endeavour was made to enhance the calculation efficiency per unit energy. As long as current computer performance is based on physical materials like silicon chips, it will eventually reach technical limits. This explains recent questions concerning whether the semiconductor industry is ready for the end of this prediction (Waldrop, 2016, Rotman, 2020).

Doing things within boundaries

This gives us the first implication: do things within boundaries, which is crucial for researchers to keep away from the 'hype' of novel methodologies or technologies and set realistic goals for their own work. The advancement of computer performance has enabled data science to flourish (Efron and Hastie, 2016b), but infinite improvement of computer performance should not be expected. Research always faces limits of various sorts, which is unavoidable. Doing things within the boundary mainly refers to finding a relatively better solution within these objective limits. For example, in medical research, very few researchers have set as a primary objective that of prolonging human life into immortality or a longevity of enormous range. But most efforts aim to defer premature death and enhance the quality of life.

In this thesis, we performed a stratified sampling approach due to the restrictions of data accessibility, which may not be optimal compared to analysing data for a study population including all patients consulting with RTI over all study years. By

acknowledging the limitations, the stratification criteria were set to be GP practices and individual study years, based on previous study findings. Supervised machine learning approaches were deployed given that patients with and without pneumonia were 'labelled' as such by clinicians. But when the quantity of data goes beyond the labelling capability of individual doctors and data at hand does not contain the information on prior knowledge of outcome values, supervised machine learning algorithms are generally unsuitable, then unsupervised machine learning approaches should be considered instead. Further, multiple variable selection methods were employed, because there is no model that performs best for all problems and fits well for all kinds of data as denoted by 'no free lunch theorem' in computer science (Wolpert and Macready, 1997). Therefore, research was performed to find a relative robust answer for the given research question rather aiming for the best answer that may not even exist.

Understanding the historical trajectory of research topic

The second implication of Moore's law is that it is important to have some knowledge of the history of the research topic. This includes the evolutionary benchmarks of a branch of science and the reasons behind them, including the bottlenecks that hampered the way and how they were eliminated. These are exactly the fundamental elements to explore. In the context of this thesis, the definition of CAP was initially developed to guide empirical antibiotic choice according to the common causal pathogens in the environment where the patients acquired the infection. Major health policies, or initiatives in antibiotic stewardship, eventually impact on coding behaviour in the healthcare system and result in the fluctuations in the apparent incidence of infectious conditions. Therefore, the chronological changes in the incidence rates of CAP together with related respiratory conditions were investigated. It was important to begin with epidemiological study of a research topic, the research results have identified the turning point of incidence trend being around 2010-2011. This suggested the existence of chronological changes in primary care which may not be only confined within respiratory conditions. And such changes are not expected to parallel with each other. Hence, prediction models generated from data of different years could be slightly different as shown by the

sensitivity analysis using subset of data from 2014 to 2017. It could be more clinically relevant for prediction modelling study using recent data as long as the sample size is sufficient even if this is at the expense of larger sample size by aggregating dated information.

Efficiency as an essential evaluation criterion

Moore's law denotes the fundamental role of efficiency of energy employed for computation. From transistor to integrated circuit, technology is set to improve the efficiency of per unit computational energy, which fits the generic rule of evolution pointing at 'energy conservation'. That is, given the same amount of resources, the evolutionary direction is always one that would yield more efficiency or save costs, if not in the short term then it- must be so in the long run. This gives us a simple rule to evaluate if a piece of research or even investment is worthwhile: only if the planned work could generate incremental benefits in terms of improvements in efficiency or saving in cost.

Prediction modelling using existing big data could be regarded as using information to exchange for energy with machine learning algorithms aiming to enhance the efficiency of such processes. In another word, prediction modelling is trying to find a simplified model to explain the majority of the question at hand. George Box's aphorism 'All models are wrong, but some are useful' denotes the essence of the information that statistical models do not fully capture the complexity of reality but still contribute the value of their utility (Box et al., 2005). The famous quotation 'Everything should be made as simple as possible, but not simpler' articulates the principle of Occam's razor which has wide applications ranging from science, biology to religion (Schaffer, 2015). In statistical modelling, the preference for simplicity could be illustrated by penalized regression models dealing with over-fitting problems. An excessively complex model is difficult to capture the underlying structure, thus may not always have better predictive performance than a simpler model. Similarly, for the same prediction, when competing hypotheses are presented, the heuristic approach is to choose the one with fewer assumptions (Anderson, 2002). In this thesis, prediction modelling results based on variable selection through

machine learning algorithms have shown that four or five variables out of more than 30 variables are able to achieve acceptable predictive performance. The models developed are able to provide simple yet robust foundations for further research to be developed, which function much the same as low magnification microscope to make a preliminary observation of a specimen. Therefore, prediction models generated from big data processed by machine learning algorithms are generally more efficient compared to embarking a new study with similar data quantity and information dimensions.

Literature review following systematic approaches as a starting point

In order to determine the incremental benefits of planned research, it is fundamental to know where we are and then we can identify where we need to be. This entails the necessity of systematic review in research. It is important to acknowledge that much of the time researchers are doing incremental rather than entirely novel work. The systematic review does not merely serve for identification of current knowledge gap but having a relative objective understanding of research area following a methodological, comprehensive fashion and reported in a transparent way, so that the research is replicable (Siddaway et al., 2019). Given that the main objectives of the systematic review should be efficient in comparison to conducting a new piece of research and critically appraising reviewed research evidence, the reviewed studies are generally set to be published and unpublished research papers rather than books. This is partially because reviewing research papers is quicker than reviewing books, more importantly determined by the difference in writing style between academic papers and books in general. It is crucial for academic researchers to disseminate research results from a relatively impartial perspective, whereas book (here does not include textbooks) authors may be prone to hold particular views like Bayesian versus frequentist. Therefore, the reviewed materials are mainly research papers from multiple databases, so that a comprehensive understanding of research topic is possible.

Systematically conducting the review following valid methodology is essential, not just for systematic review but for any type of research. Even if the results generated

from systematic approaches were similar to tossing a coin or random guess, the systematic process under methodological guidance makes the result replicable so that the authenticity, feasibility of the study could be checked and transferred to related research questions, more importantly, enables further improvement, which may make the difference.

The specific implication of Moore's law reflected in this thesis is the study of antibiotic prescription in the English general practice. Just like Moore's law which denotes the revolutionary benchmark in computer science is to increase the calculation efficiency per unit energy, antibiotics have modified the disease profile of modern history especially in infectious medicine field with pneumonia being the eminent case. Therefore, understanding antibiotic utilization in the community eventually contribute to sensible inference of prediction modelling results. Also, it served to explore possible bias of research topic since antibiotics constitute the essential element of CAP management which are not allocated to patients by randomization. Examine the key role directly associated with research question at hand will provide an important perspective to interpret study findings even if there is no apparent linkage to the main objective of the study.

Information- big data from electronic health records (EHRs)

Data structures and algorithms constitute the two fundamental elements of computer programming (Wirth, 1986). Without knowing the representation and underlying structure of the data, it is difficult to choose and construct the algorithms applied to them, vice versa, processing raw data into structured data has to reference chosen algorithms. Therefore, data structures and computer algorithms are intertwined in almost all aspects with each other.

In this thesis, understanding data structure preceded the application of algorithms. Because it is crucial to have clear objectives of data employed to answer the research question before analysis is performed on them. The main issues related to data structuring in this research include how EHR data become the repository of healthcare information as discussed in chapter three and, how these HER data were

constructed for prediction modelling which was discussed in chapter seven and eight. Here, how EHR data in primary care was bounded with CAP is discussed below.

Using existing information to update evidence on common conditions

Given the function and anatomy structure, respiratory and gastrointestinal systems are constantly exposed to the environment, which explains that most historical pandemics involved these two systems (LePan, 2020). However, with enhanced infection control and prevention procedures in modern society especially in disrupting ingestion transmission route, recent pandemics were mainly found with respiratory conditions being the common presentations such H1N1 (influenza A virus subtype H1N1) in 2009 and current Covid-19 (LePan, 2020, Huremović, 2019, Liu et al., 2020). Also, RTIs remain the leading common presenting condition in primary care in both developed and developing countries worldwide (Finley et al., 2018). It is also noticed that many research studies around CAP have presented with various results, yet similar conclusions as discussed in chapter one and six. Nevertheless, innovative research finding offering coherent comprehensive understanding of CAP is still lacking (Brown, 2012). These indicate that respiratory infection conditions are a group of common conditions accompanying human history for long term with many individual components of the conditions have been well studied. Therefore, using enriched existing data like EHR information from primary care where most CAP are managed, to update evidence of overall disease profile of subgroups of CAP patients is considered to be efficient.

Instruction commands- computer algorithms

Recursion is a central problem-solving approach for computer algorithms which makes them computationally efficient by dividing a complex problem into smaller sub-problems with the same type of their origin, then repetitive executions could be applied, and results are combined to resolve the complex problem. Recursive algorithms call themselves repetitively until the base condition or the stopping condition is met. Such a top-down branching structure approach is different from iteration, which repetitions are implemented using a looping construct until the

condition fails (CodeIT, 2020). Given that the implementation of recursion does not require prior knowledge, unlike the predefined looping construct in iteration, but solely relies on the repetitions of the algorithms, recursive approach sometimes is considered to be effective to deal with complicated problems. Whereas for iterative approach, the successful implementation needs foreknowledge or to set up a block of statements, and the looping function is repeated using output from one iteration as the input to the next. Iteration patterns are generally more familiar to humans as the natural learning process mimics the looping function- one begins with the basic knowledge and gradually upgrades to more advanced knowledge step by step. This thesis is, by no means, trying to elucidate these two approaches, rather considering the relevant implications for research.

Recursion for study design

First, the recursive approach starts from the overall problem itself rather than parts of the problem. When there is little information on the problem or the complex interactions between different parts of the problem, having a helicopter view by examining the broad aspects of the research question could efficiently harness the fundamental elements of the question. Because a research question is embedded in a context and compounded in different factors at various levels, it is difficult to manage the problem from one point in the system. As discussed previously, the data documented in EHRs is the composite reflection of the interaction between individual patients and healthcare system, as well as how EHR system functions and is deployed during healthcare service delivery at both individual and population levels. Therefore, it may not be practical to initiate the investigation of the research topic by relying on expert opinion as this may risk omitting components that were not captured by specialist knowledge.

In this thesis, antibiotic prescription in UK primary care was investigated for common infectious indications rather than pneumonia alone. Because antibiotic therapy is the fundamental or even eventual treatment for most infectious conditions managed in primary care irrespective the treatment outcome. It is difficult to have a up to date prior knowledge about how this group of drugs were being prescribed

during daily clinical practice at population level or which specific conditions might account for substantial proportions of antibiotic prescriptions in the general population. During the study, the most frequent respiratory codes for antibiotic prescription were found to be cough and chest infection. Even if this was a simple tabulation, it confirmed two points: cough is the most frequent respiratory related code, even if it is just a symptom, which could be included as an individual predictor; chest infection, which is not a formal diagnosis and easily being missed out in a code list relying on expert opinion, actually accounted for larger proportion of antibiotic prescriptions than codes suggesting substantive respiratory diagnoses like pneumonia. In the epidemiological study of pneumonia incidence trends, the incidence rate of clinically suspected pneumonia defined as antibiotic treated chest infection was at least nine times than that of clinically diagnosed pneumonia. These study findings informed prediction modelling study, also entailed the clinical relevance of developed model in the context of UK primary care system.

Objectivity vs subjectivity in machine learning

Secondly, the advantage of the recursive approach is that the fast repetition of the same algorithm until the base condition is satisfied enables computational efficiency. The base condition or in another word the ‘mission’ of the computation is generally set to minimize the difference between observed outcomes and predicted outcomes. In this prediction modelling study, this mission was set to be misclassification error during variable selection. However, the mission of the algorithms may not necessarily be the objective of the study. That is the algorithms only aim to develop accurate models rather than making sense out of data according to the study scenario, which calls for the human input in order to generate a sensible model relevant to the research question. This will generate the question of whether human input violates the objectivity of the study result?

Objectivity is a philosophical concept that contrasts with subjectivity. In science, objectivity mainly requires researchers to deliver the study process, judge the study results and disseminate the study information without partiality or under the influence of any individual entity, to assure academic rigour (Honderich, 2005). This

does not place subjectivity in an inferior place in scientific research as the most insightful interpretation to deepen the understanding of a research topic derives from subjective critical reflections (Greenhalgh et al., 2018). Also, some objective statements like ‘It either rains tomorrow, or not’ are probably the least useful observations about real life. Equally, most statistical measures or methodologies were originated from subjective conceptualization. From the choice of p value to current ML algorithms, none of them has natural existence but were developed from valid scientific theories. Therefore, elimination of subjective interpretation or discretion may diverge from the essential principle indicated by objectivity: impartial or unbiased. Research studies crafted through subjective interpretation and critique from multi-perspectives following structured approaches are still considered to be valid, including the Delphi method (Thangaratinam and Redman, 2005) and narrative review (MacLure, 2005). In ML, to generate objective results does not mean researchers should do as little as possible to the model and let the algorithms dominate the modelling process rather, introduce subjective decisions without manipulation based on evidence from various sources like medical literature and previous even interim study results. It is essentially about understanding a discipline of certain form meanwhile allowing freedom for research studies to be imposed onto that form, which does not go beyond the realms of comprehension.

In this thesis, studies from multiple perspectives deploying corresponding study designs and methodologies were implemented to serve prediction model development. For example, pneumonia was examined from its treatment (chapter four) and chronological trend (chapter five); code lists for data extraction were sorted based on expert opinion and natural language processing (chapter three); candidate predictors were sought from systematic review with multiple sources (chapter six); variable selection adopted both conventional approaches and ML algorithms (chapter seven) was implemented; model development was based on variable selection results, clinical relevance and usefulness; major bias related to final modelling results were sourced through these series studies as discussed in (from chapter three to six and chapter eight).

Iteration for study implementation

As discussed earlier before, recursion is efficient to deal with complex problems especially when little knowledge is available at the initial phase. But in order to ensure the model works better for specific research questions in later stages and at various immediate levels, refinement through an iterative process might be beneficial. A real-life example to explain this is empowerment in modern leadership. Apart from the popular virtues of empowerment like motivation, job commitment, sense of self realizations (Lee et al., 2018), empowerment is generally efficient for decision-making, especially when the size of the organization is large, or the structure of the organization is complex, or the changing pace in that industry is fast or the combination of any (Appelo, 2015). The function of such organizations depends on numerous of decisions at various levels. It is necessary that problem solving starts from the central level by capturing the majority of problem aspects other than any point from lower level. But it is almost impossible to have all the information at the central level to generate good decisions that function well everywhere across the whole organization. That explains why micromanagement is criticised because it is not efficient and if the central command does not work well, then the whole organization fails. Empowerment may allow decisions to be made from available information at local level under the constraints of central command. That is the realization of the overall objective is through iteration at local level with the foreknowledge being central command. If every level is exercising the same decision, it will become feudalistic structure; if each level is self-organising without central decisions, then the outcome may be less predictable.

The prediction modelling for the thesis started from variable selection following an iterative process as reported in chapter eight. For random forest, the initial hyperparameter (Ntree) was identified based on OOB misclassification error. Then another hyperparameter (Mtry) with array of values (5 to 30 with increment being 5) was applied based on chosen tree number (50). Then random forest models from less fitted to over fitted ones were developed to detect the most predictive variables. Candidate variables short listed based on averaged model importance ranks were selected for model development. A similar iterative process was applied to penalized

regression models. Another case is the pruning procedure for classification and regression tree (CART). CART is a typical ‘greedy’ recursive approach which only looks down to the problem in the next step. But the pruning process is iterative and factoring in considerations specific to the scenario. That is, for a given complex problem, study design could adopt recursive perspective but realized through an iterative process

Clinical significance in prediction modelling

Finally, it is crucial to incorporate clinical significance into prediction modelling process. Statistically modelling or mathematical equations sometimes could be regarded as using a quantitative language to describe the nature of research question at hand. That is, certain medical statistical models share much the same concept that could be explained by clinical knowledge which share much the same concept of wave-particle duality (Boyer et al., 1938). For example, during survival analysis where time effect of certain hazardous (or protective) factors are investigated, the analysis aims to find out if the outcome of interest resulted from the accumulation of such effects over predefined time. This would require that such effects remain to be constant within acceptable range (may not necessarily be absolutely static) over time, which is reflected by the validity of proportional hazards assumption (Bewick et al., 2004). Also, when there are alternative factors could preclude the outcome of interest being observed or reported, competing risk is considered (Zhang, 2017). Another example of such is the indication of Hagen-Poiseuille equation ($R=8\eta l/\pi r^4$, R: resistance, η : the dynamic viscosity, l: length of pipe, r: the radius of the pipe) to the determinants of airway resistance. Because of the fourth power of the radius of the pipe, a relatively small change in the radius of airway will lead to a significant increase in airway resistance even if the assumptions of the equation are not strictly held when it is applied to respiratory physiology (West, 2012). By understanding the concept, it is apparent why bronchodilators constitute the essential role in acute asthma management (NICE, 2019a). Therefore, rules or models that are valid in research generally stay valid out of research.

The ‘criticism’ of machine learning of being a ‘Black box’

One of the major criticisms that machine learning algorithms received is that they are often referred to as a ‘black box’, which points at the complexity of how risk factors behave and how the effects of individual factors contribute to the outcome of interest (Weng et al., 2019, Olden and Jackson, 2002). This, by no means, the limitation of machine learning algorithms but relies in the excessive amount of necessary background knowledge required for intended audiences or users of machine learning research outputs compared with those generated from conventional modelling approaches, which critical appraisal and judicious application are enabled. The technical explanation for this phenomenon could adopt channel capacity in information theory or bandwidth in computer science which is defined as the maximum rate of information can be reliably transmitted over a communication channel (Cover, 1999). The key implication of this definition states the channel capacity is determined by the maximum mutual information between the input and output of the channel (Cover, 1999). Therefore, information transmission efficiency largely depends on the mutual information shared between senders and receivers. A real-life example that adopts this notion is that researchers are asked to write for their readers when disseminating their research results through publications (Dixon, 2001). The subtext is that what is the general professional knowledge that the majority of the intended audience share in this field? This determines how key messages could be organized into a logical structure in order to communicate with the readers efficiently. If certain terms or jargons are common sense for both researchers and intended audiences, then a brief description should be ample otherwise a relatively granular detailed explanation is needed. This, in return, affects the communication efficiency when limited word count is applied to the draft. Another classic example in medical practice is that doctors are not allowed to talk to patients and their family using medical terms to avoid inefficient communications and issues derived from insufficient understanding due to mismatched information between two parties in such scenarios (Rimmer, 2014).

When it comes to AI in healthcare, the possible solution is to incorporate training and educational programmes into healthcare professional development curriculums at

various levels, which is advocated by health care academics recently (Matheny et al., 2020b). Because the trend that AI will influence healthcare service as well as medical research at accelerating speeds is very unlikely to be stopped or being slowed down, and bypassing the fact that equipping AI product consumers with necessary knowledge is the fundamental element to resolve the 'black box' issue, shall never lead to efficient deployment of AI technology in healthcare (Matheny et al., 2020a). Admittedly, transparently reporting machine learning modelling process, validating analytical methodology and data visualizations together with other efforts such as involving human factors (WHO, 2016a) made by researchers are able to enhance the communication efficiency of study outputs, the lack of methodology literacy of product consumers does not constitute the limitations of algorithms themselves.

In summary, the concept of using efficiency as reference criteria, following systematic approach from multiple perspectives has running through the whole thesis: from conceiving the research question, study design to conducting each individual study.

9.3 Strengths and limitations of this thesis

This section addresses the general strengths and limitations of this thesis, with detailed discussing presented in each study chapters (chapter four, five, six and eight).

The first strength relates to the overall study design of the thesis, research findings from previous studies contributed to the final prediction modelling study. Examining research question from various angles were integrated to generate a coherent and comprehensive understanding of research results of the thesis as a whole. Second, the application of both machine learning and conventional modelling approach has provided complementary information to each other. Research findings generated from machine learning algorithms have detected novel and tangible disease profile that may not be captured by traditional methodology. And to our knowledge, this is the first study applying machine learning algorithms to routine primary care EHRs to

investigate pneumonia patient profile after initial RTI consultations in a large general population. Third, all potential predictors such as frailty with information documented in the data set were explored. This would allow the final results and conclusion of the thesis to be robust. Fourth, the reporting of each studies has followed the guidelines of individual disciplines. Finally, the thesis benefits from adopting a nationally representative database of primary care EHR for analysis. Both the representativeness and quantity of research data entailed the quality of the study.

The limitations of CPRD data relating to this thesis derived from the bias introduced by inaccessibility of free texting information as discussed in chapter four. Therefore, information documented as free texting form such as patient's symptoms were not able to be investigated. Also, as illustrated in chapter eight, missing values of some predictors may carry certain meaning. Because offering and holding back certain treatment or measurements depends on various factors such as healthcare service seeking behaviour, clinical discretions or the availability of certain tests. Therefore, certain counter intuitive results were presented. Finally, the study population only captured regular resident patients presented during routine primary care consultations. Healthcare services delivered in out-of-hour, walk-in centre, nursing home, private sector, emergency and secondary care facilities were not included.

9.4 Conclusions and future implications

The main objective of this thesis was to understand the characteristics of RTI patients who presented with pneumonia within 30 days in the UK primary care settings through prediction modelling using EHRs. During prediction modelling study, machine learning model together with conventional logistic regression model were deployed. Research results demonstrated that machine learning techniques applied to routine EHR data have the potentials to identify novel disease pattern that would otherwise not have been captured using standard approaches. Future epidemiological studies then could be designed specifically based on individualized data-driven models from big data sources to verify these signals and support the initial stated research hypothesis.

References

- AABENHUS, R., HANSEN, M. P., SAUST, L. T. & BJERRUM, L. 2017. Characterisation of antibiotic prescriptions for acute respiratory tract infections in Danish general practice: a retrospective registry based cohort study. *npj Primary Care Respiratory Medicine*, 27.
- ABELLEIRA, R., RUANO-RAVINA, A., LAMA, A., BARBEITO, G., TOUBES, M. E., DOMINGUEZ-ANTELO, C., GONZALEZ-BARCALA, F. J., RODRIGUEZ-NUNEZ, N., MARCOS, P. J., PEREZ DEL MOLINO, M. L. & VALDES, L. 2019. Influenza A H1N1 Community-Acquired Pneumonia: Characteristics and Risk Factors-A Case-Control Study. *Canadian Respiratory Journal*, 2019, 4301039.
- ABERNETHY, A. P., AHMAD, A., ZAFAR, S. Y., WHEELER, J. L., REESE, J. B. & LYERLY, H. K. 2010. Electronic patient-reported data capture as a foundation of rapid learning cancer care. *Medical care*, S32-S38.
- AFONSO, A. M., EBELL, M. H., GONZALES, R., STEIN, J., GENTON, B. & SENN, N. 2012. The use of classification and regression trees to predict the likelihood of seasonal influenza. *Family Practice*, 29, 671-677.
- AGNIEL, D., KOHANE, I. S. & WEBER, G. M. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, k1479.
- AKAGI, T., NAGATA, N., MIYAZAKI, H., HARADA, T., TAKEDA, S., YOSHIDA, Y., WADA, K., FUJITA, M. & WATANABE, K. 2019. Procalcitonin is not an independent predictor of 30-day mortality, albeit predicts pneumonia severity in patients with pneumonia acquired outside the hospital. *BMC geriatrics*, 19, 3.
- AKINYEDE, O. & SOYEMI, K. 2016. Joinpoint regression analysis of pertussis crude incidence rates, Illinois, 1990-2014. *American Journal of Infection Control*, 44, 1732-1733.
- AL-GHANEM, S., AL-JAHDALI, H., BAMEFLEH, H. & KHAN, A. N. 2008. Bronchiolitis obliterans organizing pneumonia: pathogenesis, clinical features, imaging and therapy review. *Annals of thoracic medicine*, 3, 67-75.
- AL-HELOU, G., KAY, D., AHARI, J. & LESKY, L. 2016. Validation of the modified CRB-65 pneumonia severity index as a prognostic tool. *Chest*, 150 (4 Supplement 1), 1243A.

- ALBAUM, M. N., HILL, L. C., MURPHY, M., LI, Y.-H., FUHRMAN, C. R., BRITTON, C. A., KAPOOR, W. N. & FINE, M. J. 1996. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *Chest*, 110, 343-350.
- ALEXANDER, C. & WANG, L. 2017. Big data in healthcare: A New frontier in personalized medicine. *Am J Hypertens Res*, 1, 15-18.
- ALIBERTI, S., DELA, C. C., SOTGIU, G. & RESTREPO, M. I. 2019. Pneumonia is a neglected problem: it is now time to act. *The Lancet. Respiratory medicine*, 7, 10.
- ALMIRALL, J., BOLIBAR, I., SERRA-PRAT, M., PALOMERA, E., ROIG, J., HOSPITAL, I., CARANDELL, E., AGUSTI, M., AYUSO, P., ESTELA, A. & TORRES, A. 2013. Relationship between the Use of Inhaled Steroids for Chronic Respiratory Diseases and Early Outcomes in Community-Acquired Pneumonia. *PLoS ONE*, 8.
- ALMIRALL, J., SERRA-PRAT, M. & BOLIBAR, I. 2016. Risk Factors for Community-acquired Pneumonia in Adults: A Review. *Clinical Pulmonary Medicine*, 23, 99-104.
- ALMIRALL, J., SERRA-PRAT, M., BOLÍBAR, I. & BALASSO, V. 2017. Risk Factors for Community-Acquired Pneumonia in Adults: A Systematic Review of Observational Studies. *Respiration*, 94, 299-311.
- ALTMAN, D. G. 2001. Systematic reviews of evaluations of prognostic variables. *Bmj*, 323, 224-228.
- ALTMAN, D. G. 2014. Categorizing continuous variables. *Wiley StatsRef: Statistics Reference Online*.
- ALTMAN, D. G. & BLAND, J. M. 2007. Missing data. *BMJ*, 334, 424-424.
- ALTMAN, D. G. & ROYSTON, P. 2006. The cost of dichotomising continuous variables. *Bmj*, 332, 1080.
- ANALYTICS, C. 2018. EndNote X9.
- ANDERSON, D. L. 2002. Occam's razor: Simplicity, complexity, and global geodynamics. *Proceedings of the American Philosophical Society*, 146, 56-76.
- ANDREWS, J. E., RICHESSON, R. L. & KRISCHER, J. 2007. Variation of SNOMED CT coding of clinical research concepts among coding experts. *Journal of the American Medical Informatics Association*, 14, 497-506.
- ANONYMOUS 2010. Antipsychotic drugs in elderly patients associated with increased risk of pneumonia. *Australian Journal of Pharmacy*, 91, 83.

APPELO, J. 2015. *The Sense And Nonsense Of Empowerment* [Online]. Forbes. Available: <https://www.forbes.com/sites/forbes-personal-shopper/2020/08/07/what-were-buying-this-week-from-robot-window-cleaners-to-camping-cookware> [Accessed 11/08/2020].

ARCAVI, L. & BENOWITZ, N. L. 2004. Cigarette Smoking and Infection. *JAMA Internal Medicine*, 164, 2206-2216.

ARNOLD, F. W., SUMMERSGILL, J. T., LAJOIE, A. S., PEYRANI, P., MARRIE, T. J., ROSSI, P., BLASI, F., FERNANDEZ, P., FILE JR, T. M. & RELLO, J. 2007. A worldwide perspective of atypical pathogens in community-acquired pneumonia. *American journal of respiratory and critical care medicine*, 175, 1086-1093.

AROZULLAH, A. M., DALEY, J., HENDERSON, W. G., KHURI, S. F. & PROGRAM, N. V. A. S. Q. I. 2000. Multifactorial risk index for predicting postoperative respiratory failure in men after major noncardiac surgery. *Annals of surgery*, 232, 242.

ASHIRU-OREDOPE, D., SHARLAND, M., CHARANI, E., MCNULTY, C. & COOKE, J. 2012a. Improving the quality of antibiotic prescribing in the NHS by developing a new Antimicrobial Stewardship Programme: Start Smart—Then Focus. *Journal of antimicrobial chemotherapy*, 67, i51-i63.

ASHIRU-OREDOPE, D., SHARLAND, M., CHARANI, E., MCNULTY, C. & COOKE, J. 2012b. Improving the quality of antibiotic prescribing in the NHS by developing a new Antimicrobial Stewardship Programme: Start Smart—Then Focus. *Journal of antimicrobial chemotherapy*, 67, i51-i63.

ASHWORTH, M., CHARLTON, J., BALLARD, K., LATINOVIC, R. & GULLIFORD, M. 2005. Variations in antibiotic prescribing and consultation rates for acute respiratory infection in UK general practices 1995–2000. *Br J Gen Pract*, 55, 603-608.

ASHWORTH, M., CHARLTON, J., LATINOVIC, R. & GULLIFORD, M. 2006. Age-related changes in consultations and antibiotic prescribing for acute respiratory infections, 1995–2000. Data from the UK General Practice Research Database. *Journal of clinical pharmacy and therapeutics*, 31, 461-467.

ASHWORTH, M., COX, K., LATINOVIC, R., CHARLTON, J., GULLIFORD, M. & ROWLANDS, G. 2004a. Why has antibiotic prescribing for respiratory illness declined in primary care? A longitudinal study using the General Practice Research Database. *Journal of Public Health*, 26, 268-274.

ASHWORTH, M., GOLDING, S. & MAJEED, A. 2002. Prescribing indicators and their use by primary care groups to influence prescribing. *Journal of clinical pharmacy and therapeutics*, 27, 197-204.

ASHWORTH, M., LATINOVIC, R., CHARLTON, J., COX, K., ROWLANDS, G. & GULLIFORD, M. 2004b. Why has antibiotic prescribing for respiratory illness declined in primary care? A longitudinal study using the General Practice Research Database. *J Public Health (Oxf)*. 26, 268-274.

BMA. 2013. British National Formulary.

ATS, A. T. S. 1993. Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis, assessment of severity, and initial antimicrobial therapy. *Am Rev Dis*, 148, 1418-1426.

ATTIA, Z. I., NOSEWORTHY, P. A., LOPEZ-JIMENEZ, F., ASIRVATHAM, S. J., DESHMUKH, A. J., GERSH, B. J., CARTER, R. E., YAO, X., RABINSTEIN, A. A. & ERICKSON, B. J. 2019. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*.

AUSTIN, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46, 399-424.

AUSTIN, P. C. & STEYERBERG, E. W. 2017. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research*, 26, 796-808.

AUSTIN, P. C., VAN KLAVEREN, D., VERGOUWE, Y., NIEBOER, D., LEE, D. S. & STEYERBERG, E. W. 2017. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. *Diagnostic and prognostic research*, 1, 1-8.

BAIK, I., CURHAN, G. C., RIMM, E. B., BENDICH, A., WILLETT, W. C. & FAWZI, W. W. 2000. A prospective study of age and lifestyle factors in relation to community-acquired pneumonia in US men and women. *Archives of Internal Medicine*, 160, 3082-3088.

BARDSLEY, M., BLUNT, I., DAVIES, S. & DIXON, J. 2013. Is secondary preventive care improving? Observational study of 10-year trends in emergency admissions for conditions amenable to ambulatory care. *BMJ open*, 3, e002007.

BASKARAN, V., MURRAY, R. L., HUNTER, A., LIM, W. S. & MCKEEVER, T. M. 2019. Effect of tobacco smoking on the risk of developing community acquired pneumonia: A systematic review and meta-analysis. *PloS one*, 14, e0220204.

BASTIDAS, A. R., PEREZ RIVEROS, E., DELGADO GOMEZ, J., SEGURA, A. & BOTERO ROSAS, D. 2019. CURB-65 validity through use of artificial

intelligence for multiple outcomes in community acquired pneumonia. *American Journal of Respiratory and Critical Care Medicine. Conference*, 199.

BAXENDALE, H. E. & BROWN, J. S. 2012. Mechanisms of immune protection to pneumococcal infection in the young and the elderly. *Immunosenescence*. Springer.

BEAM, A. L. & KOHANE, I. S. 2018. Big Data and Machine Learning in Health Care. *JAMA*, 319, 1317-1318.

BEGLEY, C. G. & IOANNIDIS, J. P. A. 2015. Reproducibility in Science. *Circulation Research*, 116, 116-126.

BENCHIMOL, E. I., SMEETH, L., GUTTMANN, A., HARRON, K., MOHER, D., PETERSEN, I., SØRENSEN, H. T., VON ELM, E., LANGAN, S. M. & COMMITTEE, R. W. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS medicine*, 12, e1001885.

BERG, P. & LINDHARDT, B. O. 2012. The role of procalcitonin in adult patients with community-acquired pneumonia - A systematic review. *Danish Medical Journal*, 59.

BEWICK, V., CHEEK, L. & BALL, J. 2004. Statistics review 12: survival analysis. *Critical care*, 8, 389.

BHASKARAN, K., FORBES, H. J., DOUGLAS, I., LEON, D. A. & SMEETH, L. 2013. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open*, 3, e003389.

BLUMENTALS, W. A., NEVITT, A., PENG, M. M. & TOOVEY, S. 2012. Body mass index and the incidence of influenza-associated pneumonia in a UK primary care cohort. *Influenza and other Respiratory Viruses*, 6, 28-36.

BLUMENTHAL, D. & DIXON, J. 2012. Health-care reforms in the USA and England: areas for useful learning. *The Lancet*, 380, 1352-1357.

BLUNT, I. 2013. Focus on preventable admissions. *London: Nuffield Trust*.

BMA, B. M. A. 2013. British National Formulary (65). 351-355.

BOERSMA, E., POLDERMANS, D., BAX, J. J., STEYERBERG, E. W., THOMSON, I. R., BANGA, J. D., VAN DE VEN, L. L., VAN URK, H., ROELANDT, J. R. & GROUP, D. S. 2001. Predictors of cardiac events after major vascular surgery: role of clinical characteristics, dobutamine echocardiography, and β -blocker therapy. *Jama*, 285, 1865-1873.

- BONT, J., HAK, E., HOES, A., SCHIPPER, M., SCHELLEVIS, F. & VERHEIJ, T. 2007. A prediction rule for elderly primary-care patients with lower respiratory tract infections. *European Respiratory Journal*, 29, 969-975.
- BONTEN, M. J., HUIJTS, S. M., BOLKENBAAS, M., WEBBER, C., PATTERSON, S., GAULT, S., VAN WERKHOVEN, C. H., VAN DEURSEN, A. M., SANDERS, E. A. & VERHEIJ, T. J. 2015. Polysaccharide conjugate vaccine against pneumococcal pneumonia in adults. *New England Journal of Medicine*, 372, 1114-1125.
- BOOTH, H. P., PREVOST, A. T. & GULLIFORD, M. C. 2013. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. *Pharmacoepidemiology and drug safety*, 22, 1357-1361.
- BOOTH, H. P., PREVOST, A. T. & GULLIFORD, M. C. 2015. Access to weight reduction interventions for overweight and obese patients in UK primary care: population-based cohort study. *BMJ open*, 5, e006642.
- BOUSTANI, M. A., MUNGER, S., GULATI, R., VOGEL, M., BECK, R. A. & CALLAHAN, C. M. 2010. Selecting a change and evaluating its impact on the performance of a complex adaptive health care delivery system. *Clinical interventions in aging*, 5, 141.
- BOX, G. E. P., HUNTER, J. S., HUNTER, W. G., BINS, R., KIRLIN IV, K. & CARROLL, D. 2005. *Statistics for experimenters: design, innovation, and discovery*, Wiley New York.
- BOYD, A., THOMAS, R. & MACLEOD, J. 2018. NHS Number and the Systems Used to Manage Them: An Overview for Research Users. *London: CLOSER*.
- BOYER, C., EINSTEIN, A. & INFELD, L. 1938. *The Evolution of Physics: The Growth of Ideas From Early Concepts to Relativity and Quanta*. Cambridge: Cambridge University Press (2nd edition, 1961).
- BRADLEY, J. S., BYINGTON, C. L., SHAH, S. S., ALVERSON, B., CARTER, E. R., HARRISON, C., KAPLAN, S. L., MACE, S. E., MCCracken JR, G. H. & MOORE, M. R. 2011. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the Pediatric Infectious Diseases Society and the Infectious Diseases Society of America. *Clinical infectious diseases*, 53, e25-e76.
- BREIMAN, L. 2001. Random forests. *Machine learning*, 45, 5-32.
- BRINSLEY, K., SRINIVASAN, A., SINKOWITZ-COCHRAN, R., LAWTON, R., MCINTYRE, R., KRAVITZ, G., BURKE, B., SHADOWEN, R. & CARDO, D.

2005. Implementation of the campaign to prevent antimicrobial resistance in healthcare settings: 12 steps to prevent antimicrobial resistance among hospitalized adults—experiences from 3 institutions. *American journal of infection control*, 33, 53-54.
- BROWN, J. S. 2012. Community-acquired pneumonia. *Clinical Medicine*, 12, 538.
- BTS. 2015. 2015 - Annotated BTS Guideline for the management of CAP in adults (2009) Summary of recommendations.
- BUENDIA, F., BALANAG, V. & CECILIA JOCSON, M. A. 2010. The use of CURB-65 as a severity assessment tool for community acquired pneumonia among patients at the lung center of the philippines. *Respirology*, 15, 44.
- BURTON, C. L., CHESTERTON, L. S., CHEN, Y. & VAN DER WINDT, D. A. 2016. Clinical course and prognostic factors in conservatively managed carpal tunnel syndrome: a systematic review. *Archives of physical medicine and rehabilitation*, 97, 836-852. e1.
- CALDEIRA, D., ALARCÃO, J., VAZ-CARNEIRO, A. & COSTA, J. 2012. Risk of pneumonia associated with use of angiotensin converting enzyme inhibitors and angiotensin receptor blockers: systematic review and meta-analysis. *Bmj*, 345, e4260.
- CAMPBELL, D. T. & COOK, T. D. 1979. *Quasi-experimentation: Design & analysis issues for field settings*, Rand McNally College Publishing Company Chicago.
- CANET, J., GALLART, L., GOMAR, C., PALUZIE, G., VALLES, J., CASTILLO, J., SABATE, S., MAZO, V., BRIONES, Z. & SANCHIS, J. 2010. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 113, 1338-1350.
- CANTO, J. G., KIEFE, C. I., ROGERS, W. J., PETERSON, E. D., FREDERICK, P. D., FRENCH, W. J., GIBSON, C. M., POLLACK, C. V., ORNATO, J. P. & ZALENSKI, R. J. 2011. Number of coronary heart disease risk factors and mortality in patients with first myocardial infarction. *Jama*, 306, 2120-2127.
- CARBON, C. & BAX, R. P. 1998. Regulating the use of antibiotics in the community. *BMJ: British Medical Journal*, 317, 663.
- CARRIQUIRY, A., HOFMANN, H., TAI, X. H. & VANDERPLAS, S. 2019. Machine learning in forensic applications. *Significance*, 16, 29-35.

CDC. 2017. *Pneumococcal Vaccination: What Everyone Should Know* [Online]. Available: <https://www.cdc.gov/vaccines/vpd/pneumo/public/index.html> [Accessed 16/10/2019 2019].

CDC. 2020. *International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)* [Online]. Available: <https://www.cdc.gov/nchs/icd/icd10cm.htm> [Accessed 10/01/2020].

CHALMERS, J., CAMPLING, J., ELLSBURY, G., HAWKEY, P. M., MADHAVA, H. & SLACK, M. 2017. Community-acquired pneumonia in the United Kingdom: a call to action. *Pneumonia*, 9, 15.

CHALMERS, J. D., SINGANAYAGAM, A., AKRAM, A. R., MANDAL, P., SHORT, P. M., CHOUDHURY, G., WOOD, V. & HILL, A. T. 2010. Severity assessment tools for predicting mortality in hospitalised patients with community-acquired pneumonia. Systematic review and meta-analysis. *Thorax*, 65, 878-883.

CHAN, M. 2009. *World now at the start of 2009 influenza pandemic* [Online]. World Health Organisation. Available: https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/ [Accessed 28st November 2018].

CHARLSON, M. E., POMPEI, P., ALES, K. L. & MACKENZIE, C. R. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40, 373-383.

CHATENOUD, L., GARAVELLO, W., PAGAN, E., BERTUCCIO, P., GALLUS, S., LA VECCHIA, C., NEGRI, E. & BOSETTI, C. 2016. Laryngeal cancer mortality trends in European countries. *International Journal of Cancer*, 138, 833-842.

CHATTERJEE, S., CARNAHAN, R. M., CHEN, H., HOLMES, H. M., JOHNSON, M. L. & APARASU, R. R. 2016. Anticholinergic medication use and risk of pneumonia in elderly adults: A nested case-control study. *Journal of the American Geriatrics Society*, 64, 394-400.

CHAURASIA, A. R. 2020a. COVID-19 Trend and Forecast in India: A Joinpoint Regression Analysis. *medRxiv*.

CHAURASIA, A. R. 2020b. Long-term trend in infant mortality in India: a joinpoint regression analysis for 1981-2018. *medRxiv*.

CHEN, J. H., ALAGAPPAN, M., GOLDSTEIN, M. K., ASCH, S. M. & ALTMAN, R. B. 2017. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *International journal of medical informatics*, 102, 71-79.

CHEN, J. H. & ASCH, S. M. 2017. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine*, 376, 2507.

CHERRYMAN, G. 2006. Imaging in primary care. *British Journal of General Practice*, 56, 563-564.

CHICCO, D. & JURMAN, G. 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20, 16.

CHICCO, D. & ROVELLI, C. 2019. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PloS one*, 14, e0208737.

CHO, W., BYUN, K., KANG, L., JEON, D. & KIM, Y. 2017. Prognostic significance of nutritional risk in the elderly patients with community acquired pneumonia. *American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS*, 195.

CHOBY, B. A. & HUNTER, P. 2015. Respiratory infections: community-acquired pneumonia. *Fp Essentials*, 429, 11-21.

CHOI, S.-H., KIM, E. Y. & KIM, Y.-J. 2013. Systemic use of fluoroquinolone in children. *Korean journal of pediatrics*, 56, 196-201.

CICORIA, S., SHERLOCK, J., MUNISWAMIAIAH, M. & CLARKE, L. Classification of titanic passenger data and chances of surviving the disaster. Proceedings of Student-Faculty Research Day, CSIS, 2014. 1-6.

CILLONIZ, C., POLVERINO, E., EWIG, S., ALIBERTI, S., GABARRUS, A., MENENDEZ, R., MENSA, J., BLASI, F. & TORRES, A. 2013. Impact of age and comorbidity on cause and outcome in community-acquired pneumonia. *Chest*, 144, 999-1007.

CLARK, K., GOLDSTEIN, R. L., HART, J. E., TEYLAN, M., LAZZARI, A. A., GAGNON, D. R., TUN, C. G. & GARSHICK, E. 2020. Plasma vitamin D, past chest illness, and risk of future chest illness in chronic spinal cord injury (SCI): a longitudinal observational study. *Spinal Cord*, 16, 16.

CLEGG, A., BATES, C., YOUNG, J., RYAN, R., NICHOLS, L., ANN TEALE, E., MOHAMMED, M. A., PARRY, J. & MARSHALL, T. 2016. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and ageing*, 45, 353-360.

CLEGG, A., YOUNG, J., ILIFFE, S., RIKKERT, M. O. & ROCKWOOD, K. 2013. Frailty in elderly people. *The lancet*, 381, 752-762.

CODEIT. 2020. *Differences between iterative and recursive algorithms* [Online]. Erasmus. Available: <https://www.codeit-project.eu/differences-between-iterative-and-recursive-algorithms/> [Accessed].

COHEN, S. M., LEE, H. J., LEIMAN, D. A., ROY, N. & MISONO, S. 2019. Associations between Community-Acquired Pneumonia and Proton Pump Inhibitors in the Laryngeal/Voice-Disordered Population. *Otolaryngology - Head and Neck Surgery (United States)*, 161, 546-547.

COLLINS, G. S. & MOONS, K. G. 2019. Reporting of artificial intelligence prediction models. *The Lancet*, 393, 1577-1579.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC medicine*, 13, 1.

COMMITTEE, J. F. 2017. BNF 74: September 2017. London, UK: Pharmaceutical Press.

COMMITTEE, S. M. A. The path of least resistance. London: Department of Health, 1998.

CONTROL, C. F. D. & PREVENTION 2013. *Antibiotic resistance threats in the United States, 2013*, Centres for Disease Control and Prevention, US Department of Health and Human Services.

COOKE, J., BUTLER, C., HOPSTAKEN, R., DRYDEN, M. S., MCNULTY, C., HURDING, S., MOORE, M. & LIVERMORE, D. M. 2015. Narrative review of primary care point-of-care testing (POCT) and antibacterial use in respiratory tract infection (RTI). *BMJ Open Respiratory Research*, 2, e000086.

CORRALES-MEDINA, V. F., MADJID, M. & MUSER, D. M. 2010. Role of acute infection in triggering acute coronary syndromes. *The Lancet infectious diseases*, 10, 83-92.

CORRALES-MEDINA, V. F., MUSER, D. M., SHACHKINA, S. & CHIRINOS, J. A. 2013. Acute pneumonia and the cardiovascular system. *The Lancet*, 381, 496-505.

CORRALES-MEDINA, V. F., MUSER, D. M., WELLS, G. A., CHIRINOS, J. A., CHEN, L. & FINE, M. J. 2012. Cardiac complications in patients with community-acquired pneumonia: incidence, timing, risk factors, and association with short-term mortality. *Circulation*, 125, 773-781.

- COSTELLOE, C., METCALFE, C., LOVERING, A., MANT, D. & HAY, A. D. 2010. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *Bmj*, 340, c2096.
- COVER, T. M. 1999. *Elements of information theory*, John Wiley & Sons.
- CPRD. 2019a. *CPRD linked data* [Online]. Available: <https://www.cprd.com/linked-data> [Accessed 17/12/2019].
- CPRD. 2019b. *Primary care data for public health research* [Online]. Available: <https://www.cprd.com/primary-care> [Accessed 17/12/2019].
- CPRD. 2021. *CPRD linked data* [Online]. Available: <https://cprd.com/linked-data#Patient%20postcode%20linked%20measures> [Accessed 14/02 2021].
- CRD. 2009. *Systematic reviews: CRD's guidance for undertaking reviews in health care*, Centre for Reviews and Dissemination.
- CUNHA, B. 2006. The atypical pneumonias: clinical diagnosis and importance. *Clinical Microbiology and Infection*, 12, 12-24.
- D RILEY, R., A VAN DER WINDT, D., CROFT, P. & GM MOONS, K. 2019. Fundamental statistical methods for prognosis research. *Prognosis research in Healthcare: Concepts, Methods and Impact*. 1 ed.: Oxford University Press.
- DAABISS, M. 2011. American Society of Anaesthesiologists physical status classification. *Indian journal of anaesthesia*, 55, 111.
- DANG, T. T., MAJUMDAR, S. R., MARRIE, T. J. & EURICH, D. T. 2015. Recurrent pneumonia: a review with focus on clinical epidemiology and modifiable risk factors in elderly patients. *Drugs & Aging*, 32, 13-9.
- DAROUICHE, R., MOSIER, M. & VOIGT, J. 2012. Antibiotics and antiseptics to prevent infection in cardiac rhythm management device implantation surgery. *Pacing and Clinical Electrophysiology*, 35, 1348-1360.
- DAVIES, S. & GIBBENS, N. 2013. UK five year antimicrobial resistance strategy 2013 to 2018. *London: Department of Health*.
- DAVIES, S. C., FOWLER, T., WATSON, J., LIVERMORE, D. M. & WALKER, D. 2013. Annual Report of the Chief Medical Officer: infection and the rise of antimicrobial resistance. *The Lancet*, 381, 1606-1609.
- DEAN, N. C., GRIFFITH, P. P., SORENSEN, J. S., MCCAULEY, L., JONES, B. E. & LEE, Y. G. 2016. Pleural effusions at first ED encounter predict worse clinical outcomes in patients with pneumonia. *Chest*, 149, 1509-1515.

- DEENY, S. R. & STEVENTON, A. 2015. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Quality & Safety*, 24, 505-515.
- DELMESTRI, A. & PRIETO-ALHAMBRA, D. 2020. CPRD GOLD and linked ONS mortality records: Reconciling guidelines. *International journal of medical informatics*, 136, 104038.
- DEMIR, A. Y. 2014. Implementation of c-reactive protein (CRP) point of care testing in the primary care: A pilot study. *Clinical Chemistry and Laboratory Medicine*, 52, S1521.
- DENT, E., MARTIN, F. C., BERGMAN, H., WOO, J., ROMERO-ORTUNO, R. & WALSTON, J. D. 2019. Management of frailty: opportunities, challenges, and future directions. *The Lancet*, 394, 1376-1386.
- DIXON, N. 2001. Writing for publication—a guide for new authors. *International Journal for Quality in Health Care*, 13, 417-421.
- DOLK, F. C. K., POUWELS, K. B., SMITH, D. R. M., ROBOTHAM, J. V. & SMIESZEK, T. 2018. Antibiotics in primary care in England: which antibiotics are prescribed and for which conditions? *Journal of Antimicrobial Chemotherapy*, 73, ii2-ii10.
- DONG, P. & CREMER, O. 2011. Limitations of the use of the Glasgow Coma Scale in intensive care patients with non-neurological primary disease: a search for alternatives. *Critical Care*, 15, P506.
- DOSHI, S. M., RUEDA, A. M., CORRALES-MEDINA, V. F. & MUSER, D. M. 2011. Anemia and community-acquired pneumococcal pneumonia. *Infection*, 39, 379-383.
- DOUCET, M., ROCHETTE, L. & HAMEL, D. 2016. Incidence, prevalence, and mortality trends in chronic obstructive pulmonary disease over 2001 to 2011: a public health point of view of the burden. *Canadian Respiratory Journal*, 2016.
- DRIJKONINGEN, J. & ROHDE, G. 2014. Pneumococcal infection in adults: burden of disease. *Clinical Microbiology and Infection*, 20, 45-51.
- DUBLIN, S., WALKER, R. L., JACKSON, M. L., NELSON, J. C., WEISS, N. S. & JACKSON, L. A. 2011. Use of angiotensin-converting enzyme inhibitors is not associated with decreased pneumonia risk. *Pharmacoepidemiology and Drug Safety*, 20, S83-S84.
- DUBLIN, S., WALKER, R. L., JACKSON, M. L., NELSON, J. C., WEISS, N. S. & JACKSON, L. A. 2012. Angiotensin-converting enzyme inhibitor use and

pneumonia risk in community-dwelling older adults: Results from a population-based case-control study. *Pharmacoepidemiology and Drug Safety*, 21, 1173-1182.

EASTERBROOK, P. J., GOPALAN, R., BERLIN, J. & MATTHEWS, D. R. 1991. Publication bias in clinical research. *The Lancet*, 337, 867-872.

ECCLES, R. 2002. An explanation for the seasonality of acute upper respiratory tract viral infections. *Acta oto-laryngologica*, 122, 183-191.

ECCLES, S., PINCUS, C., HIGGINS, B. & WOODHEAD, M. 2014. Diagnosis and management of community and hospital acquired pneumonia in adults: summary of NICE guidance. *Bmj*, 349, g6722.

EDELMAN, E. J., GORDON, K. S., CROTHERS, K., AKGUN, K., BRYANT, K. J., BECKER, W. C., GAITHER, J. R., GIBERT, C. L., GORDON, A. J., MARSHALL, B. D. L., RODRIGUEZ-BARRADAS, M. C., SAMET, J. H., JUSTICE, A. C., TATE, J. P. & FIELLIN, D. A. 2019. Association of Prescribed Opioids with Increased Risk of Community-Acquired Pneumonia among Patients with and Without HIV. *JAMA Internal Medicine*, 179, 297-304.

EFRON, B. & HASTIE, T. 2016a. Computer age statistical inference. Cambridge University Press.

EFRON, B. & HASTIE, T. 2016b. *Computer age statistical inference*, Cambridge University Press.

EIZADI-MOOD, N., MAZROEI-SEBEDANI, S., SOLTANINEJAD, F. & BABAK, A. 2018. Risk factors associated with aspiration pneumonia among the patients with drug intoxication. *Journal of isfahan medical school*, 36, 510-516.

EKELUND, S. 2012. ROC curves—What are they and how are they used? *Point of Care*, 11, 16-21.

EOM, C. S., JEON, C. Y., LIM, J. W., CHO, E. G., PARK, S. M. & LEE, K. S. 2011. Use of acid-suppressive drugs and risk of pneumonia: A systematic review and meta-analysis. *Cmaj*, 183, 310-319.

EPLER, G. R. 2001. Bronchiolitis obliterans organizing pneumonia. *Archives of Internal Medicine*, 161, 158-164.

ESR & WONCA, E. 2010. Radiology and primary care in Europe. *Insights into Imaging*, 1, 46.

EURICH, D. T., LEE, C., MARRIE, T. J. & MAJUMDAR, S. R. 2013. Inhaled corticosteroids and risk of recurrent pneumonia: A population-based, nested case-control study. *Clinical Infectious Diseases*, 57, 1138-1144.

- EURICH, D. T., MARRIE, T. J., MINHAS-SANDHU, J. K. & MAJUMDAR, S. R. 2017. Risk of heart failure after community acquired pneumonia: prospective controlled study with 10 years of follow-up. *bmj*, 356, j413.
- EWIG, S., WELTE, T., CHASTRE, J. & TORRES, A. 2010. Rethinking the concepts of community-acquired and health-care-associated pneumonia. *The Lancet infectious diseases*, 10, 279-287.
- FACULTY OF HEALTH SCIENCES, T. U. O. S. 2018. *General Practice Classifications and Terminologies* [Online]. Available: <https://sydney.edu.au/health-sciences/ncch/icpc-2-plus/overview.shtml> [Accessed 10/01/2020].
- FALCONE, M., VENDITTI, M., SHINDO, Y. & KOLLEF, M. H. 2011. Healthcare-associated pneumonia: diagnostic criteria and distinction from community-acquired pneumonia. *International Journal of Infectious Diseases*, 15, e545-e550.
- FANG, L., KARAKIULAKIS, G. & ROTH, M. 2020. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet. Respiratory Medicine*, 8, e21.
- FAVERIO, P. & SIBILA, O. 2017. New biomarkers in community-acquired pneumonia: Another step in improving outcome prediction. *Respirology*, 22, 416-417.
- FIELDING, J. E. 1985. Smoking: health effects and control. *New England journal of medicine*, 313, 491-498.
- FILE JR, T. M. & MARRIE, T. J. 2010. Burden of community-acquired pneumonia in North American adults. *Postgraduate medicine*, 122, 130-141.
- FINE, M. J., AUBLE, T. E., YEALY, D. M., HANUSA, B. H., WEISSFELD, L. A., SINGER, D. E., COLEY, C. M., MARRIE, T. J. & KAPOOR, W. N. 1997. A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia. *New England Journal of Medicine*, 336, 243-250.
- FINLEY, C. R., CHAN, D. S., GARRISON, S., KOROWNYK, C., KOLBER, M. R., CAMPBELL, S., EURICH, D. T., LINDBLAD, A. J., VANDERMEER, B. & ALLAN, G. M. 2018. What are the most common conditions in primary care?: Systematic review. *Canadian Family Physician*, 64, 832-840.
- FLEGAL, K. M., GRAUBARD, B. I., WILLIAMSON, D. F. & GAIL, M. H. 2005. Excess deaths associated with underweight, overweight, and obesity. *Jama*, 293, 1861-1867.
- FLETCHER, R. H., FLETCHER, S. W. & FLETCHER, G. S. 2012. *Clinical epidemiology: the essentials*, Lippincott Williams & Wilkins.

FLORKOWSKI, C. M. 2008. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29, S83.

FOX, J. & WEISBERG, S. 2019. *An R Companion to Applied Regression*. Thousand Oaks CA: Sage.

FRANKLIN, M. & THORN, J. 2019. Self-reported and routinely collected electronic healthcare resource-use data for trial-based economic evaluations: the current state of play in England and considerations for the future. *BMC medical research methodology*, 19, 8.

FRICK, J., MÖCKEL, M., MULLER, R., SEARLE, J., SOMASUNDARAM, R. & SLAGMAN, A. 2017. Suitability of current definitions of ambulatory care sensitive conditions for research in emergency department patients: a secondary health data analysis. *BMJ open*, 7, e016109.

FRIED, L. P., TANGEN, C. M., WALSTON, J., NEWMAN, A. B., HIRSCH, C., GOTTDIENER, J., SEEMAN, T., TRACY, R., KOP, W. J. & BURKE, G. 2001. Frailty in older adults: evidence for a phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56, M146-M157.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent.

FRISCHER, M., HEATLIE, H., NORWOOD, J., BASHFORD, J., MILLSON, D. & CHAPMAN, S. 2001. Trends in antibiotic prescribing and associated indications in primary care from 1993 to 1997. *Journal of Public Health*, 23, 69-73.

FRY, A. M., SHAY, D. K., HOLMAN, R. C., CURNS, A. T. & ANDERSON, L. J. 2005. Trends in hospitalizations for pneumonia among persons aged 65 years or older in the United States, 1988-2002. *Jama*, 294, 2712-2719.

FUNG, H. B. & MONTEAGUDO-CHU, M. O. 2010. Community-acquired pneumonia in the elderly. *American Journal Geriatric Pharmacotherapy*, 8, 47-62.

GARCIA GARRIDO, H. M., MAK, A. M. R., WIT, F. W. N. M., WONG, G. W. M., KNOL, M. J., VOLLAARD, A., TANCK, M. W. T., VAN DER ENDE, A., GROBUSCH, M. P. & GOORHUIS, A. 2019. Incidence and Risk Factors for Invasive Pneumococcal Disease and Community-acquired Pneumonia in Human Immunodeficiency Virus-Infected Individuals in a High-income Setting. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 21.

- GARIN, N., MARTI, C., SCHEFFLER, M., STIRNEMANN, J. & PRENDKI, V. 2019. Computed tomography scan contribution to the diagnosis of community-acquired pneumonia. *Current opinion in pulmonary medicine*, 25, 242-248.
- GAYNES, R. 2017. The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerging Infectious Diseases*, 23, 849.
- GBD 2015 LRI COLLABORATORS 2017. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17, 1133-1161.
- GEORGE, B., SEALS, S. & ABAN, I. 2014. Survival analysis and regression models. *Journal of Nuclear Cardiology*, 21, 686-694.
- GIULIANO, C., WILHELM, S. M. & KALE-PRADHAN, P. B. 2012. Are proton pump inhibitors associated with the development of community-acquired pneumonia? A meta-analysis. *Expert Review of Clinical Pharmacology*, 5, 337-44.
- GLICKSBERG, B. S., JOHNSON, K. W. & DUDLEY, J. T. 2018. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Human Molecular Genetics*, 27, R56-R62.
- GONZALEZ DEL CASTILLO, J., CLEMENTE, C., NUNEZ ORANTOS, M. J. & EN REPRESENTACION DE, I.-S. 2019. Risk stratification of patients with pneumonia. *Medicina Clinica*, 152, e21.
- GOOSSENS, H., FERECHE, M., VANDER STICHELE, R., ELSEVIERS, M. & GROUP, E. P. 2005. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *The Lancet*, 365, 579-587.
- GREENHALGH, T., THORNE, S. & MALTERUD, K. 2018. Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, 48, e12931.
- GREENWELL, B., BOEHMKE, B. & GRAY, B. 2020. vip: Variable Importance Plots.
- GRIJALVA, C. G., ZHU, Y., NUORTI, J. P. & GRIFFIN, M. R. 2011. Emergence of parapneumonic empyema in the USA. *Thorax*, 66, 663-668.
- GROENWOLD, R. H., DONDERS, A. R. T., ROES, K. C., HARRELL JR, F. E. & MOONS, K. G. 2011. Dealing with missing outcome data in randomized trials and observational studies. *American journal of epidemiology*, 175, 210-217.

GROENWOLD, R. H., DONDERS, A. R. T., ROES, K. C., HARRELL JR, F. E. & MOONS, K. G. 2012a. Dealing with missing outcome data in randomized trials and observational studies. *American journal of epidemiology*, 175, 210-217.

GROENWOLD, R. H., WHITE, I. R., DONDERS, A. R. T., CARPENTER, J. R., ALTMAN, D. G. & MOONS, K. G. 2012b. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Cmaj*, 184, 1265-1269.

GROSIOS, K., GAHAN, P. B. & BURBIDGE, J. 2010. Overview of healthcare in the UK. *EPMA Journal*, 1, 529-534.

GULLIFORD, M. 2019. Frailty and Drug Safety at Older Ages. *Drug safety*, 42, 699-700.

GULLIFORD, M., LATINOVIC, R., CHARLTON, J., LITTLE, P., VAN STAA, T. & ASHWORTH, M. 2009a. Selective decrease in consultations and antibiotic prescribing for acute respiratory tract infections in UK primary care up to 2006. *Journal of public health*, 31, 512-520.

GULLIFORD, M. C., CHARLTON, J., ASHWORTH, M., RUDD, A. G., TOSCHKE, A. M. & TEAM, E. R. 2009b. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PLoS One*, 4, e7168.

GULLIFORD, M. C., CHARLTON, J., WINTER, J. R., SUN, X., REZEL-POTTS, E., BUNCE, C., FOX, R., LITTLE, P., HAY, A. D. & MOORE, M. V. 2020. Probability of sepsis after infection consultations in primary care in the United Kingdom in 2002–2017: Population-based cohort study and decision analytic model. *PLoS Medicine*, 17, e1003202.

GULLIFORD, M. C., DREGAN, A., MOORE, M. V., ASHWORTH, M., STAA, T., MCCANN, G., CHARLTON, J., YARDLEY, L., LITTLE, P. & MCDERMOTT, L. 2014a. Continued high rates of antibiotic prescribing to adults with respiratory tract infection: survey of 568 UK general practices. *BMJ Open*, 4, e006245.

GULLIFORD, M. C., DREGAN, A., MOORE, M. V., ASHWORTH, M., VAN STAA, T., MCCANN, G., CHARLTON, J., YARDLEY, L., LITTLE, P. & MCDERMOTT, L. 2014b. Continued high rates of antibiotic prescribing to adults with respiratory tract infection: survey of 568 UK general practices. *BMJ open*, 4, e006245.

GULLIFORD, M. C., MOORE, M. V., LITTLE, P., HAY, A. D., FOX, R., PREVOST, A. T., JUSZCZYK, D., CHARLTON, J. & ASHWORTH, M. 2016. Safety of reduced antibiotic prescribing for self limiting respiratory tract infections in primary care: cohort study using electronic health records. *bmj*, 354, i3410.

- HADDA, V., MADAN, M., MITTAL, S., MADAN, K. & ESQUINAS, A. 2019. Severe community acquired pneumonia: Prediction of outcome. *Journal of critical care*, 54, 287.
- HAGGERTY, C. M., JAMES, C. A., CALKINS, H., TICHNELL, C., LEADER, J. B., HARTZEL, D. N., NEVIUS, C. D., PENDERGRASS, S. A., PERSON, T. N. & SCHWARTZ, M. 2017. Electronic health record phenotype in subjects with genetic variants associated with arrhythmogenic right ventricular cardiomyopathy: a study of 30,716 subjects with exome sequencing. *Genetics in Medicine*, 19, 1245.
- HAM, J. Y. & EUN SONG, K. 2019. A prospective study of presepsin as an indicator of the severity of community-acquired pneumonia in emergency departments: Comparison with pneumonia severity index and CURB-65 scores. *Lab Medicine*, 50, 364-369.
- HANLON, P., NICHOLL, B. I., JANI, B. D., LEE, D., MCQUEENIE, R. & MAIR, F. S. 2018. Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: a prospective analysis of 493 737 UK Biobank participants. *The Lancet Public Health*, 3, e323-e332.
- HARRELL JR, F. E. 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, Springer.
- HARSHFIELD, A., RHODES, K., BARCLAY, S. & PAYNE, R. 2017. The accuracy of death dates recorded in the clinical practice research datalink (CPRD).
- HAWKER, J. I., SMITH, S., SMITH, G. E., MORBEY, R., JOHNSON, A. P., FLEMING, D. M., SHALLCROSS, L. & HAYWARD, A. C. 2014. Trends in antibiotic prescribing in primary care for clinical syndromes subject to national recommendations to reduce antibiotic resistance, UK 1995–2011: analysis of a large database of primary care consultations. *Journal of Antimicrobial Chemotherapy*, 69, 3423-3430.
- HAWN, M. T., RICHMAN, J. S., VICK, C. C., DEIERHOI, R. J., GRAHAM, L. A., HENDERSON, W. G. & ITANI, K. M. 2013. Timing of surgical antibiotic prophylaxis and the risk of surgical site infection. *JAMA surgery*, 148, 649-657.
- HAYDEN, J. A., CHOU, R., HOGG-JOHNSON, S. & BOMBARDIER, C. 2009. Systematic reviews of low back pain prognosis had variable methods and results: guidance for future prognosis reviews. *Journal of clinical epidemiology*, 62, 781-796.e1.
- HAYDEN, J. A., CÔTÉ, P. & BOMBARDIER, C. 2006. Evaluation of the quality of prognosis studies in systematic reviews. *Annals of internal medicine*, 144, 427-437.

HAYDEN, J. A., VAN DER WINDT, D. A., CARTWRIGHT, J. L., CÔTÉ, P. & BOMBARDIER, C. 2013. Assessing bias in studies of prognostic factors. *Annals of internal medicine*, 158, 280-286.

HÄYRINEN, K., SARANTO, K. & NYKÄNEN, P. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77, 291-304.

HE, F., WU, X., XIA, X., PENG, F., HUANG, F. & YU, X. 2013. Pneumonia and mortality risk in continuous ambulatory peritoneal dialysis patients with diabetic nephropathy. *PloS one*, 8.

HE, Z., GELLER, J. & CHEN, Y. 2015. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. *Artificial intelligence in medicine*, 64, 29-40.

HEAD, M. G., FITCHETT, J. R., COOKE, M. K., WURIE, F. B., HAYWARD, A. C., LIPMAN, M. C. & ATUN, R. 2014. Investments in respiratory infectious disease research 1997–2010: a systematic analysis of UK funding. *BMJ Open*, 4, e004600.

HEALE, R. & FORBES, D. 2013. Understanding triangulation in research. *Evidence Based Nursing*, 16, 98-98.

HEATH, A., GONZALES, M. & VON ALVENSLEBEN, I. 2019. Variable selection for early diagnosis of congenital heart disease using random forest entropy calculations. *Cardiology in the Young*, 29 (Supplement 1), S187.

HELD, U., BOVE, D. S., STEURER, J. & HELD, L. 2012. Validating and updating a risk model for pneumonia - a case study. *BMC medical research methodology*, 12, 99.

HELLEN GELBAND, M. M.-P., SURAJ PANT, SUMANTH GANDRA, JORDAN LEVINSON, DEVRA BARTER, ANDREA WHITE, RAMANAN LAXMINARAYAN. 2015. *The State of the World's Antibiotics, 2015* [Online]. Center for disease dynamics, economics policy. Available: http://www.cddep.org/publications/state_worlds_antibiotics_2015#sthash.2fHwn4BD.dpbs [Accessed 30th August 2016].

HEMINGWAY, H., CROFT, P., PEREL, P., HAYDEN, J. A., ABRAMS, K., TIMMIS, A., BRIGGS, A., UDUMYAN, R., MOONS, K. G. & STEYERBERG, E. W. 2013. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *Bmj*, 346, e5595.

HEMINGWAY, H., RILEY, R. D. & ALTMAN, D. G. 2009. Ten steps towards improving prognosis research. *BMJ*, 339, b4184.

HERRETT, E., GALLAGHER, A. M., BHASKARAN, K., FORBES, H., MATHUR, R., VAN STAA, T. & SMEETH, L. 2015. Data resource profile: clinical practice research datalink (CPRD). *International journal of epidemiology*, 44, 827-836.

HERRETT, E., THOMAS, S. L., SCHOONEN, W. M., SMEETH, L. & HALL, A. J. 2010. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *British journal of clinical pharmacology*, 69, 4-14.

HESS, G., HILL, J. W., RAUT, M. K., FISHER, A. C., MODY, S., SCHEIN, J. R. & CHEN, C. C. 2010. Comparative antibiotic failure rates in the treatment of community-acquired pneumonia: Results from a claims analysis. *Advances in Therapy*, 27, 743-755.

HICKEY, G. L., KONTOPANTELIS, E., TAKKENBERG, J. J. M. & BEYERSDORF, F. 2018. Statistical primer: checking model assumptions with regression diagnostics†. *Interactive CardioVascular and Thoracic Surgery*, 28, 1-8.

HINGORANI, A. D., VAN DER WINDT, D. A., RILEY, R. D., ABRAMS, K., MOONS, K. G., STEYERBERG, E. W., SCHROTER, S., SAUERBREI, W., ALTMAN, D. G. & HEMINGWAY, H. 2013. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *Bmj*, 346, e5793.

HIPPOCRATES & ADAMS, F. 1939. *The Genuine Works of Hippocrates; Translated from the Greek by Francis Adams*, Bailliere, Tindall & Cox.

HOFFMANN, T. J., EHRET, G. B., NANDAKUMAR, P., RANATUNGA, D., SCHAEFER, C., KWOK, P.-Y., IRIBARREN, C., CHAKRAVARTI, A. & RISCH, N. 2017. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature genetics*, 49, 54.

HONDERICH, T. 2005. *The Oxford companion to philosophy*, OUP Oxford.

HONEYFORD, K., COOKE, G. S., KINDERLERER, A., WILLIAMSON, E., GILCHRIST, M., HOLMES, A., ROOM, T. S. B., GLAMPSON, B., MULLA, A. & COSTELLOE, C. 2019. Evaluating a digital sepsis alert in a London multisite hospital network: a natural experiment using electronic health record data. *Journal of the American Medical Informatics Association*.

HOOGENDIJK, E. O., AFILALO, J., ENSRUD, K. E., KOWAL, P., ONDER, G. & FRIED, L. P. 2019. Frailty: implications for clinical practice and public health. *The Lancet*, 394, 1365-1375.

HOOGENDIJK, E. O., DEEG, D. J. H., POPPELAARS, J., VAN DER HORST, M., BROESE VAN GROENOU, M. I., COMIJS, H. C., PASMAN, H. R. W., VAN SCHOOR, N. M., SUANET, B., THOMÉSE, F., VAN TILBURG, T. G., VISSER,

- M. & HUISMAN, M. 2016. The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. *European Journal of Epidemiology*, 31, 927-945.
- HOPSTAKEN, R., WITBRAAD, T., VAN ENGELSHOVEN, J. & DINANT, G. 2004. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. *Clinical radiology*, 59, 743-752.
- HORIE, M., UGAJIN, M., SUZUKI, M., NOGUCHI, S., TANAKA, W., YOSHIHARA, H., KAWAKAMI, M., KICHIKAWA, Y. & SAKAMOTO, Y. 2012. Diagnostic and prognostic value of procalcitonin in community-acquired pneumonia. *American Journal of the Medical Sciences*, 343, 30-35.
- HORNE, B. D., MAY, H. T., MUHLESTEIN, J. B., RONNOW, B. S., LAPPÉ, D. L., RENLUND, D. G., KFOURY, A. G., CARLQUIST, J. F., FISHER, P. W. & PEARSON, R. R. 2009. Exceptional mortality prediction by risk scores from common laboratory tests. *The American journal of medicine*, 122, 550-558.
- HOTHORN, T., HORNIK, K. & ZEILEIS, A. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15, 651-674.
- HOWARD, L., SILLIS, M., PASTEUR, M., KAMATH, A. & HARRISON, B. 2005. Microbiological profile of community-acquired pneumonia in adults over the last 20 years. *Journal of Infection*, 50, 107-113.
- HRIPCSAK, G. & ALBERS, D. J. 2012. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20, 117-121.
- HRIPCSAK, G., KNIRSCH, C., ZHOU, L., WILCOX, A. & MELTON, G. B. 2011. Bias associated with mining electronic health records. *Journal of biomedical discovery and collaboration*, 6, 48.
- HRIPCSAK, G., KUPERMAN, G. J. & FRIEDMAN, C. 1998. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of information in medicine*, 37, 01-07.
- HRIPCSAK, G., ZHOU, L., PARSONS, S., DAS, A. K. & JOHNSON, S. B. 2005. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. *Journal of the American Medical Informatics Association*, 12, 55-63.
- HSCIC, H. S. C. I. C. 2015. *SCCI0021: ICD-10 5th Edition: Change Paper* [Online]. Available: <http://content.digital.nhs.uk/media/18526/0021102014change-spec/pdf/0021102014change-spec.pdf> [Accessed].

HUDDY, J. R., NI, M. Z., BARLOW, J., MAJEED, A. & HANNA, G. B. 2016. Point-of-care C reactive protein for the diagnosis of lower respiratory tract infection in NHS primary care: a qualitative study of barriers and facilitators to adoption. *BMJ Open*, 6, e009959.

HUILGOL, P. 2020. *Bias and Variance in Machine Learning – A Fantastic Guide for Beginners!* [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/> [Accessed 15/08/2020].

HULLEY, S., CUMMINGS, S., BROWNER, W., GRADY, D. & NEWMAN, T. 2001. Designing clinical research. Lippincott Williams & Wilkins. *Philadelphia, PA*, 97-99.

HUREMOVIĆ, D. 2019. Brief History of Pandemics (Pandemics Throughout History). *Psychiatry of Pandemics*. Springer.

HUSSEY, E. 1999. 5 Heraclitus. *The Cambridge Companion to Early Greek Philosophy*, 88.

IOANNIDIS, J. P. 2005. Why most published research findings are false. *PLoS medicine*, 2, e124.

ISHIFUJI, T., SANDO, E., KANEKO, N., SUZUKI, M., YAEGASHI, M., AOSHIMA, M., ARIYOSHI, K. & MORIMOTO, K. 2015. Medications associated with the incidence of recurrent pneumonia in Japanese elderly population. *European Respiratory Journal. Conference: European Respiratory Society Annual Congress*, 46.

JAIN, S., SELF, W. H., WUNDERINK, R. G., FAKHRAN, S., BALK, R., BRAMLEY, A. M., REED, C., GRIJALVA, C. G., ANDERSON, E. J. & COURTNEY, D. M. 2015a. Community-acquired pneumonia requiring hospitalization among US adults. *New England Journal of Medicine*, 373, 415-427.

JAIN, S., WILLIAMS, D. J., ARNOLD, S. R., AMPOFO, K., BRAMLEY, A. M., REED, C., STOCKMANN, C., ANDERSON, E. J., GRIJALVA, C. G. & SELF, W. H. 2015b. Community-acquired pneumonia requiring hospitalization among US children. *New England Journal of Medicine*, 372, 835-845.

JANSON, C., JOHANSSON, G., STÄLLBERG, B., LISSPERS, K., OLSSON, P., KEININGER, D. L., UHDE, M., GUTZWILLER, F. S., JÖRGENSEN, L. & LARSSON, K. 2018. Identifying the associated risks of pneumonia in COPD patients: ARCTIC an observational study. *Respiratory research*, 19, 172.

- JENSEN, P. B., JENSEN, L. J. & BRUNAK, S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13, 395.
- JHA, A. K., DOOLAN, D., GRANDT, D., SCOTT, T. & BATES, D. W. 2008. The use of health information technology in seven nations. *International journal of medical informatics*, 77, 848-854.
- JICK, H., TERRIS, B. Z., DERBY, L. E. & JICK, S. S. 1992. Further validation of information recorded on a general practitioner based computerized data resource in the United Kingdom. *Pharmacoepidemiology and drug safety*, 1, 347-349.
- JOHN, U. & HANKE, M. 2015. Liver cirrhosis mortality, alcohol consumption and tobacco consumption over a 62 year period in a high alcohol consumption country: a trend analysis. *BMC research notes*, 8, 822.
- JOHNSON, K. W., SHAMEER, K., GLICKSBERG, B. S., READHEAD, B., SENGUPTA, P. P., BJÖRKEGREN, J. L., KOVACIC, J. C. & DUDLEY, J. T. 2017. Enabling precision cardiology through multiscale biology and systems medicine. *JACC: Basic to Translational Science*, 2, 311-327.
- JOHNSON, O. A., FRASER, H. S., WYATT, J. C. & WALLEY, J. D. 2014. Electronic health records in the UK and USA. *The Lancet*, 384, 954.
- JUNIUS-WALKER, U., ONDER, G., SOLEYMANI, D., WIESE, B., ALBAINA, O., BERNABEI, R. & MARZETTI, E. 2018. The essence of frailty: A systematic review and qualitative synthesis on frailty concepts and definitions. *European journal of internal medicine*, 56, 3-10.
- KAELBER, D. C. & BATES, D. W. 2007. Health information exchange and patient safety. *Journal of biomedical informatics*, 40, S40-S45.
- KALRA, L., SMITH, C. J., HODSOLL, J., VAIL, A., IRSHAD, S. & MANAWADU, D. 2019. Elevated C-reactive protein increases diagnostic accuracy of algorithm-defined stroke-associated pneumonia in afebrile patients. *International journal of stroke*, 14, 167-173.
- KANNEL, W. B., MCGEE, D. & GORDON, T. 1976. A general cardiovascular risk profile: the Framingham Study. *The American journal of cardiology*, 38, 46-51.
- KAPLAN, V., ANGUS, D. C., GRIFFIN, M. F., CLERMONT, G., SCOTT WATSON, R. & LINDE-ZWIRBLE, W. T. 2002. Hospitalized community-acquired pneumonia in the elderly: age-and sex-related patterns of care and outcome in the United States. *American journal of respiratory and critical care medicine*, 165, 766-772.

KARACA-MANDIC, P., NORTON, E. C. & DOWD, B. 2012. Interaction Terms in Nonlinear Models. *Health Services Research*, 47, 255-274.

KARAKIOULAKI, M. & STOLZ, D. 2019a. Biomarkers and clinical scoring systems in community-acquired pneumonia. *Annals of Thoracic Medicine*, 14, 165-172.

KARAKIOULAKI, M. & STOLZ, D. 2019b. Biomarkers in pneumonia-beyond procalcitonin. *International Journal of Molecular Sciences*, 20.

KEW, K. M. & SENIUKOVICH, A. 2014. Inhaled steroids and risk of pneumonia for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, 2014.

KHAN, A. R., RAHBI, H., BINABDULHAK, A. A., RIAZ, M., ALTANNIR, M. A., KASHOUR, T., GARBATI, M., RAHBI, B., IBRAHIM, T. & TLEYJEH, I. M. 2011a. The association between statin use and the outcome of community acquired pneumonia: A systematic review and meta-analysis. *American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS*, 183.

KHAN, A. R., RAHBI, H., BINABDULHAK, A. A., RIAZ, M., ALTANNIR, M. A., KASHOUR, T., GARBATI, M., RAHBI, B., IBRAHIM, T. & TLEYJEH, I. M. 2011b. The association between statin use and the risk of community acquired pneumonia: A systematic review and meta-analysis. *American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS*, 183.

KHAN, N. F., PERERA, R., HARPER, S. & ROSE, P. W. 2010. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. *BMC family practice*, 11, 1.

KIM, H. J., FAY, M. P., FEUER, E. J. & MIDTHUNE, D. N. 2000. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in medicine*, 19, 335-351.

KIM, H. J., FAY, M. P., YU, B., BARRETT, M. J. & FEUER, E. J. 2004. Comparability of segmented line regression models. *Biometrics*, 60, 1005-1014.

KOJIMA, G., ILIFFE, S. & WALTERS, K. 2017. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age and ageing*, 47, 193-200.

KOMOROWSKI, M., CELI, L. A., BADAWI, O., GORDON, A. C. & FAISAL, A. A. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24, 1716.

- KONTOPANTELIS, E., DORAN, T., SPRINGATE, D. A., BUCHAN, I. & REEVES, D. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *bmj*, 350, h2750.
- KONTOPANTELIS, E., STEVENS, R. J., HELMS, P. J., EDWARDS, D., DORAN, T. & ASHCROFT, D. M. 2018. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open*, 8, e020738.
- KONTOPANTELIS, E., WHITE, I. R., SPERRIN, M. & BUCHAN, I. 2017. Outcome-sensitive multiple imputation: a simulation study. *BMC medical research methodology*, 17, 1-13.
- KOTSIANTIS, S. B., ZAHARAKIS, I. & PINTELAS, P. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.
- KOUSOULIS, A. A., RAFI, I. & DE LUSIGNAN, S. 2015. The CPRD and the RCGP: building on research success by enhancing benefits for patients and practices. *British Journal of General Practice*.
- KRENKE, K., URBANKOWSKA, E., URBANKOWSKI, T., LANGE, J. & KULUS, M. 2016. Clinical characteristics of 323 children with parapneumonic pleural effusion and pleural empyema due to community acquired pneumonia. *Journal of Infection and Chemotherapy*, 22, 292-297.
- KUHN, M. 2008. Building predictive models in R using the caret package. *Journal of statistical software*, 28, 1-26.
- KUHN, M. 2020. caret: Classification and Regression Training.
- KUHN, M. & JOHNSON, K. 2013. *Applied predictive modeling*, Springer.
- KURU, T. & LYNCH III, J. P. 1999. Nonresolving or slowly resolving pneumonia. *Clinics in chest medicine*, 20, 623-651.
- LAXMINARAYAN, R., DUSE, A., WATTAL, C., ZAIDI, A. K., WERTHEIM, H. F., SUMPRADIT, N., VLIEGHE, E., HARA, G. L., GOULD, I. M. & GOOSSENS, H. 2013. Antibiotic resistance—the need for global solutions. *The Lancet infectious diseases*, 13, 1057-1098.
- LAZER, D., KENNEDY, R., KING, G. & VESPIGNANI, A. 2014. The parable of Google Flu: traps in big data analysis. *Science*, 343, 1203-1205.
- LEE, A., WILLIS, S. & TIAN, W. A. 2018. *When Empowering Employees Works, and When It Doesn't* [Online]. Harvard Business Review. Available:

<https://hbr.org/2018/03/when-empowering-employees-works-and-when-it-doesnt>
[Accessed 15/08/2020].

LEE, J. & SONG, J. U. 2018. Pneumococcal urinary antigen test use as a prognostic marker in patients admitted with community-acquired pneumonia: A propensity score matching study. *Respirology*, 23 (Supplement 2), 307.

LEITE, A., THOMAS, S. L. & ANDREWS, N. J. 2017. Implementing near real-time vaccine safety surveillance using the Clinical Practice Research Datalink (CPRD). *Vaccine*, 35, 6885-6892.

LEPAN, N. 2020. Visualizing the history of pandemics. *Visualizing the History of Pandemics*.

LERA, R., FERNANDEZ-FABRELLAS, E., CERVERA, A., BLANQUER, J., SANZ, F., CHINER, E., SANCHO, J. N., SENENT, C. & AGUAR, M. 2011. Differential features of Community Acquired Pneumonia (CAP) in young adults. *American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS*, 183.

LEVY, M. L., LE JEUNE, I., WOODHEAD, M. A., MACFARLANE, J. T. & LIM, W. S. 2010. Primary care summary of the British Thoracic Society Guidelines for the management of community acquired pneumonia in adults: 2009 update Endorsed by the Royal College of General Practitioners and the Primary Care Respiratory Society UK. *Primary Care Respiratory Journal*, 19, 21.

LEWIS, M. & SWIRSKY, D. 2001. Spleen: consequences of lack of function. *e LS*.

LI, W., DING, C. & YIN, S. 2015. Severe pneumonia in the elderly: A multivariate analysis of risk factors. *International Journal of Clinical and Experimental Medicine*, 8, 12463-12475.

LIANG, S.-F., SUN, X., GULLIFORD, M. & CURCIN, V. Inclusion and Exclusion of Medical Codes for Primary Care Data Extraction. 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018. IEEE, 394-395.

LIAPIKOU, A., MALLIOU, I., PETRAS, P., KASIOLA, M., ANASTASOPOULOS, A., VOLONAKIS, M., CHAIMALA, D. & DIMAKOU, K. 2012. Clinical presentation and evolution of community acquired pneumonia in older patients. *European Respiratory Journal. Conference: European Respiratory Society Annual Congress*, 40.

LIAW, A. & WIENER, M. 2002. Classification and regression by randomForest. *R news*, 2, 18-22.

- LIEBERMAN, D., SHVARTZMAN, P., KORSONSKY, I. & LIEBERMAN, D. 2003. Diagnosis of ambulatory community-acquired pneumonia. *Scandinavian journal of primary health care*, 21, 57-60.
- LIM, W., MACFARLANE, J., BOSWELL, T., HARRISON, T., ROSE, D., LEINONEN, M. & SAIKKU, P. 2001a. Study of community acquired pneumonia aetiology (SCAPA) in adults admitted to hospital: implications for management guidelines. *Thorax*, 56, 296-301.
- LIM, W. S. 2015. Who is Funding What in the Fight Against Pneumonia? *EBioMedicine*, 2, 1025.
- LIM, W. S., BAUDOUIN, S., GEORGE, R., HILL, A., JAMIESON, C., LE JEUNE, I., MACFARLANE, J., READ, R., ROBERTS, H. & LEVY, M. 2009a. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. *Thorax*, 64, iii1-iii55.
- LIM, W. S., BAUDOUIN SR, S., GEORGE, R., HILL, A., JAMIESON, C., LE JEUNE, I., MACFARLANE, J., READ, R., ROBERTS, H., LEVY, M., WANI, M. & WOODHEAD, M. 2009b. British Thoracic Society guidelines for the management of community acquired pneumonia in adults: Update 2009. *Thorax*, 64, iii1-iii55.
- LIM, W. S., MACFARLANE, J. T., BOSWELL, T. C. J., HARRISON, T. G., ROSE, D., LEINONEN, M. & SAIKKU, P. 2001b. Study of community acquired pneumonia aetiology (SCAPA) in adults admitted to hospital: implications for management guidelines. *Thorax*, 56, 296-301.
- LIM, W. S., SMITH, D. L., WISE, M. P. & WELHAM, S. A. 2015. British Thoracic Society community acquired pneumonia guideline and the NICE pneumonia guideline: how they fit together. *BMJ Open Respiratory Research*, 2, e000091.
- LIM, W. S. & WOODHEAD, M. 2011. British Thoracic Society adult community acquired pneumonia audit 2009/10. *Thorax*, 66, 548-549.
- LIN, S. H., PERNG, D. W., CHEN, C. P., CHAI, W. H., YEH, C. S., KOR, C. T., CHENG, S. L., CHEN, J. J. & LIN, C. H. 2016. Increased risk of community-acquired pneumonia in COPD patients with comorbid cardiovascular disease. *International Journal of COPD*, 11, 3051-3058.
- LIN, W.-L., MUO, C.-S., LIN, W.-C., HSIEH, Y.-W. & KAO, C.-H. 2019. Association of Increased Risk of Pneumonia and Using Proton Pump Inhibitors in Patients With Type II Diabetes Mellitus. *Dose-Response*, 17, 1559325819843383.
- LINDENAUER, P. K., LAGU, T., SHIEH, M.-S., PEKOW, P. S. & ROTHBERG, M. B. 2012. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. *Jama*, 307, 1405-1413.

LITTLE, P., STUART, B., SMITH, S., THOMPSON, M. J., KNOX, K., VAN DEN BRUEL, A., LOWN, M., MOORE, M. & MANT, D. 2017. Antibiotic prescription strategies and adverse outcome for uncomplicated lower respiratory tract infections: prospective cough complication cohort (3C) study. *bmj*, 357, j2148.

LIU, Y.-C., KUO, R.-L. & SHIH, S.-R. 2020. COVID-19: The first documented coronavirus pandemic in history. *Biomedical journal*.

LOH, T., LAI, F., TAN, E., THOON, K., TEE, N., CUTTER, J. & TANG, J. 2011. Correlations between clinical illness, respiratory virus infections and climate factors in a tropical paediatric population. *Epidemiology & Infection*, 139, 1884-1894.

LONG, K. S., COULSON, E., SARAVANAN, V., HEYCOCK, C., HAMILTON, J. & KELLY, C. 2010. What factors predict pneumonia in patients with rheumatoid arthritis? *Rheumatology*, 49, i166-i167.

LU, D., ZHANG, J., MA, C., YUE, Y., ZOU, Z., YU, C. & YIN, F. 2018. Link between community-acquired pneumonia and vitamin D levels in older patients. *Zeitschrift fur Gerontologie und Geriatrie*, 51, 435-439.

MACDOUGALL, C. & POLK, R. E. 2005. Antimicrobial stewardship programs in health care systems. *Clinical microbiology reviews*, 18, 638-656.

MACKENZIE, G. 2016. The definition and classification of pneumonia. *Pneumonia*, 8, 14.

MACLURE, M. 2005. MacLure, Maggie, "Clarity Bordering on Stupidity!: Where's the Quality in Systemic Review," *Journal of Educational Policy*, 20 (No. 4, 2005), 393-416. Reprinted pp. 45-70 in Bridget Somekh and Thomas A. Schwandt, eds., *Knowledge Production: Research Work in Interesting Times*. New York: Routledge, 2007.

MAHON, C., WALKER, W., DRAGE, A. & BEST, E. 2016. Incidence, aetiology and outcome of pleural empyema and parapneumonic effusion from 1998 to 2012 in a population of New Zealand children. *Journal of Paediatrics and Child Health*, 52, 662-668.

MALEK, F., GOHARI, A., MIRMOHAMMADKHANI, M. & ARDIANI, F. 2019. Relationship between the serum level of C-reactive protein and severity and outcomes of community-acquired pneumonia. *Archives of Clinical Infectious Diseases*, 14.

MANDELL, L. A., WUNDERINK, R. G., ANZUETO, A., BARTLETT, J. G., CAMPBELL, G. D., DEAN, N. C., DOWELL, S. F., FILE JR, T. M., MUSER, D. M. & NIEDERMAN, M. S. 2007. Infectious Diseases Society of America/American

Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical infectious diseases*, 44, S27-S72.

MANGINI, S., PIRES, P. V., BRAGA, F. G. M. & BACAL, F. 2013. Decompensated heart failure. *Einstein*, 11, 383.

MAPEL, D., SCHUM, M., YOOD, M., BROWN, J., MILLER, D. & DAVIS, K. 2010. Pneumonia among COPD patients using inhaled corticosteroids and long-acting bronchodilators. *Primary Care Respiratory Journal*, 19, 109-117.

MASKELL, N. 2010. British thoracic society pleural disease guidelines-2010 update. BMJ Publishing Group Ltd.

MATHENY, M., ISRANI, S. T., AHMED, M. & WHICHER, D. 2020a. Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Natl Acad Med*, 94-97.

MATHENY, M. E., WHICHER, D. & THADANEY ISRANI, S. 2020b. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *JAMA*, 323, 509-510.

MCCABE, C., KIRCHNER, C., ZHANG, H., DALEY, J. & FISMAN, D. N. 2009. Guideline-concordant therapy and reduced mortality and length of stay in adults with community-acquired pneumonia: playing by the rules. *Archives of internal medicine*, 169, 1525-1531.

MCDONAGH, M., PETERSON, K., WINTHROP, K., CANTOR, A., HOLZHAMMER, B. & BUCKLEY, D. I. 2016. Improving antibiotic prescribing for uncomplicated acute respiratory tract infections.

METERSKY, M. L., MASTERTON, R. G., LODE, H., FILE, T. M. & BABINCHAK, T. 2012. Epidemiology, microbiology, and treatment considerations for bacterial pneumonia complicating influenza. *International Journal of Infectious Diseases*, 16, E321-E331.

METLAY, J. P. & FINE, M. J. 2003. Testing Strategies in the Initial Management of Patients with Community-Acquired Pneumonia. *Annals of Internal Medicine*, 138, 109-118.

METLAY, J. P., STAFFORD, R. S. & SINGER, D. E. 1998. National trends in the use of antibiotics by primary care physicians for adult patients with cough. *Archives of internal medicine*, 158, 1813-1818.

METLAY, J. P., WATERER, G. W., LONG, A. C., ANZUETO, A., BROZEK, J., CROTHERS, K., COOLEY, L. A., DEAN, N. C., FINE, M. J. & FLANDERS, S. A. 2019. Diagnosis and treatment of adults with community-acquired pneumonia. An

official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. *American journal of respiratory and critical care medicine*, 200, e45-e67.

METOFFICE. *When does spring start?* [Online]. Available: <https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/spring/when-does-spring-start> [Accessed 10/10/2019].

MILLETT, E. R., DE STAVOLA, B. L., QUINT, J. K., SMEETH, L. & THOMAS, S. L. 2015. Risk factors for hospital admission in the 28 days following a community-acquired pneumonia diagnosis in older adults, and their contribution to increasing hospitalisation rates over time: a cohort study. *BMJ Open*, 5, e008737.

MILLETT, E. R., QUINT, J. K., SMEETH, L., DANIEL, R. M. & THOMAS, S. L. 2013. Incidence of community-acquired lower respiratory tract infections and pneumonia among older adults in the United Kingdom: a population-based study. *PloS one*, 8, e75131.

MINNAARD, M. C., DE GROOT, J. A. H., HOPSTAKEN, R. M., SCHIERENBERG, A., DE WIT, N. J., REITSMA, J. B., BROEKHUIZEN, B. D. L., VAN VUGT, S. F., NEVEN, A. K., GRAFFELMAN, A. W., MELBYE, H., RAINER, T. H., STEURER, J., HOLM, A., GONZALES, R., DINANT, G. J., VAN DE POL, A. C. & VERHEIJ, T. J. M. 2017. The added value of C-reactive protein measurement in diagnosing pneumonia in primary care: A meta-analysis of individual patient data. *Cmaj*, 189, E56-E63.

MITCHELL, T. M. 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45, 870-877.

MIZGERD, J. P. 2008. Acute lower respiratory tract infection. *New England Journal of Medicine*, 358, 716-727.

MOHER, D., LIBERATI, A., TETZLAFF, J. & ALTMAN, D. G. 2010. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg*, 8, 336-341.

MONTGOMERY, D. C. 2020. *Introduction to Statistical quality control*, Wiley.

MOONS, K. G., DE GROOT, J. A., BOUWMEESTER, W., VERGOUWE, Y., MALLETT, S., ALTMAN, D. G., REITSMA, J. B. & COLLINS, G. S. 2014. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS medicine*, 11, e1001744.

MOONS, K. G., ROYSTON, P., VERGOUWE, Y., GROBBEE, D. E. & ALTMAN, D. G. 2009. Prognosis and prognostic research: what, why, and how? *Bmj*, 338, b375.

- MOORE, G. E. 1965. Cramming more components onto integrated circuits. McGraw-Hill New York, NY, USA:.
- MORILLO, D. S., JIMENEZ, A. L. & MORENO, S. A. 2013. Computer-aided diagnosis of pneumonia in patients with chronic obstructive pulmonary disease. *Journal of the American Medical Informatics Association*, 20, e111-e117.
- MORTENSEN, E. M., HALM, E. A., PUGH, M. J., COPELAND, L. A., METERSKY, M., FINE, M. J., JOHNSON, C. S., ALVAREZ, C. A., FREI, C. R. & GOOD, C. 2014. Association of azithromycin with mortality and cardiovascular events among older patients hospitalized with pneumonia. *Jama*, 311, 2199-2208.
- MUKAMAL, K. J., GHIMIRE, S., PANDEY, R., O'MEARA, E. S. & GAUTAM, S. 2010. Antihypertensive medications and risk of community-acquired pneumonia. *Journal of Hypertension*, 28, 401-5.
- MUSHER, D. M. & THORNER, A. R. 2014. Community-acquired pneumonia. *New England Journal of Medicine*, 371, 1619-1628.
- NAKAJIMA, M., UMEZAKI, Y., TAKEDA, S., YAMAGUCHI, M., SUZUKI, N., YONEDA, M., HIROFUJI, T., SEKITANI, H., YAMASHITA, Y. & MORITA, H. 2020. Association between oral candidiasis and bacterial pneumonia: A retrospective study. *Oral Diseases*, 26, 234-237.
- NAKANISHI, M., YOSHIDA, Y., TAKEDA, N., HIRANA, H., HORITA, T., SHIMIZU, K., HIRATANI, K., TOYODA, S., MATSUMURA, T., SHINNO, E., KAWAI, S., FUTAMURA, A., OTA, M. & NATAZUKA, T. 2010. Significance of the progression of respiratory symptoms for predicting community-acquired pneumonia in general practice. *Respirology*, 15, 969-974.
- NATHAN, C. & CARS, O. 2014. Antibiotic resistance—problems, progress, and prospects. *New England Journal of Medicine*, 371, 1761-1763.
- NATIONAL INSTITUTE FOR, H. & CLINICAL, E. 2008. *Prescribing of antibiotics for self-limiting respiratory tract infections in adults and children in primary care. Final scope 130907*, London, National Institute for Health and Clinical Excellence.
- NEMOTO, M., NAKASHIMA, K., MATSUE, Y., ISHIFUJI, T., KATSURADA, N., MORIMOTO, K., ARIYOSHI, K. & AOSHIMA, M. 2014. Incremental prognostic predict ability of chest computed tomography in patients with community onset pneumonia. *Respirology*, 19, 47.
- NHS DIGITAL. 2021. *Quality and Outcome Framework* [Online]. Available: <https://qof.digital.nhs.uk> [Accessed 15/02 2021].

- NHS DIGITAL, N. H. S. 2015. *UK Read Code* [Online]. Available: <https://data.gov.uk/dataset/f262aa32-9c4e-44f1-99eb-4900deada7a4/uk-read-code> [Accessed 10/01/2020].
- NHS DIGITAL, N. H. S. 2019a. READ V2 TO SNOMED CT MAPPING LOOKUP (October 2019).
- NHS DIGITAL, N. H. S. 2019b. *SNOMED CT* [Online]. Available: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct> [Accessed 16th May 2019].
- NHS DIGITAL, N. H. S. 2019c. *SNOMED CT implementation in primary care* [Online]. Available: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care> [Accessed 13/01/2020].
- NHS. 2016. *Notes on GP IT Systems (August 2016)* [Online]. Available: <https://www.england.nhs.uk/ourwork/accessibleinfo/resources/gp-it-systems/> [Accessed 17/12/2019].
- NHS & BMA. 2015a.
- NHS & BMA. 2015b. Seasonal influenza vaccination programme Read Codes used for payment.
- NHS. 2019a. *Flu vaccine overview* [Online]. Available: <https://www.nhs.uk/conditions/vaccinations/flu-influenza-vaccine/> [Accessed 16/10/2019 2019].
- NHS. 2019b. *NHS vaccinations and when to have them* [Online]. Available: <https://www.nhs.uk/conditions/vaccinations/nhs-vaccinations-and-when-to-have-them/> [Accessed].
- NHS. 2019c. *Pneumococcal vaccine overview* [Online]. Available: <https://www.nhs.uk/conditions/vaccinations/pneumococcal-vaccination/> [Accessed 16/10/2019 2019].
- NHS RESEARCH, S. 2020. *Electronic Health Records* [Online]. Available: <https://www.nhsresearchscotland.org.uk/research-in-scotland/data/sub-page-4> [Accessed 08/01/2020].
- NICE & EXCELLENCE, N. I. F. H. A. C. 2014. Pneumonia in adults: diagnosis and management.

NICE. 2008a. Respiratory tract infections-antibiotic prescribing: prescribing of antibiotics for self-limiting respiratory tract infections in adults and children in primary care.

NICE. 2014. Pneumonia in adults: diagnosis and management.

NICE. 2015. Chest infections - adult.

NICE. 2019a. Asthma, acute.

NICE. 2019b. Cough (acute): antimicrobial prescribing.

NICE. 2019c. *Respiratory infections: All NICE products on respiratory infections. Includes any guidance, advice, NICE Pathways and quality standards.* [Online]. Available: <https://www.nice.org.uk/guidance/conditions-and-diseases/respiratory-conditions/respiratory-infections> [Accessed 07/11 2019].

NICE. 2008b. Prescribing of antibiotics for self-limiting respiratory tract infections in adults and children in primary care. Final scope 130907. Available: <http://www.nice.org.uk/nicemedia/pdf/RTIFinalScope.pdf>.

NIE, W., ZHANG, Y., JEE, S. H., JUNG, K. J., LI, B. & XIU, Q. 2014. Obesity survival paradox in pneumonia: A meta-analysis. *BMC Medicine*, 12.

NIEDERMAN, M. S. 1998. Community-acquired pneumonia: a North American perspective. *Chest*, 113, 179S-182S.

NIEDERMAN, M. S., MANDELL, L. A., ANZUETO, A., BASS, J. B., BROUGHTON, W. A., CAMPBELL, G. D., DEAN, N., FILE, T., FINE, M. J. & GROSS, P. A. 2001. Guidelines for the management of adults with community-acquired pneumonia: diagnosis, assessment of severity, antimicrobial therapy, and prevention. *American journal of respiratory and critical care medicine*, 163, 1730-1754.

NIELSEN, A. B., THORSEN-MEYER, H.-C., BELLING, K., NIELSEN, A. P., THOMAS, C. E., CHMURA, P. J., LADEMANN, M., MOSELEY, P. L., HEIMANN, M. & DYBDAHL, L. 2019. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *The Lancet Digital Health*, 1, e78-e89.

NIH. 2018. *Number of Joints* [Online]. Available: <https://surveillance.cancer.gov/help/jointpoint/setting-parameters/method-and-parameters-tab/number-of-joints> [Accessed 18/08/2020].

- NOSE, M., RECLA, E., TRIFIRO, G. & BARBUI, C. 2015. Antipsychotic drug exposure and risk of pneumonia: A systematic review and meta-analysis of observational studies. *Pharmacoepidemiology and Drug Safety*, 24, 812-820.
- NOWAKOWSKA, M., ZGHEBI, S. S., ASHCROFT, D. M., BUCHAN, I., CHEW-GRAHAM, C., HOLT, T., MALLEEN, C., VAN MARWIJK, H., PEEK, N. & PERERA-SALAZAR, R. 2019. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. *BMC medicine*, 17, 145.
- OCHOA-GONDAR, O., VILA-CÓRCOLES, A., DE DIEGO, C., ARIJA, V., MAXENCHS, M., GRIVE, M., MARTIN, E., PINYOL, J. L. & GROUP, E.-S. 2008. The burden of community-acquired pneumonia in the elderly: the Spanish EVAN-65 study. *BMC public health*, 8, 222.
- OCHOA-GONDAR, O., VILA-CORCOLES, A., RODRIGUEZ-BLANCO, T., GOMEZ-BERTOMEU, F., FIGUEROLA-MASSANA, E., RAGA-LURIA, X. & HOSPITAL-GUARDIOLA, I. 2014. Effectiveness of the 23-valent pneumococcal polysaccharide vaccine against community-acquired pneumonia in the general population aged ≥ 60 years: 3 years of follow-up in the CAPAMIS study. *Clinical Infectious Diseases*, 58, 909-917.
- OLDEN, J. D. & JACKSON, D. A. 2002. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154, 135-150.
- OLSON, G. & DAVIS, A. M. 2020. Diagnosis and Treatment of Adults With Community-Acquired Pneumonia. *JAMA*.
- ÖRTQVIST, Å. 2001. Pneumococcal vaccination: current and future issues. *European Respiratory Journal*, 18, 184-195.
- OZLEK, E., BITEKER, F. S., CIL, C., CELIK, O., OZLEK, B., DOGAN, V. & BITEKER, M. 2019. The risk stratification in community-acquired pneumonia. *American Journal of Emergency Medicine*, 37, 171.
- PAGLIARI, C., DETMER, D. & SINGLETON, P. 2007. Potential of electronic personal health records. *Bmj*, 335, 330-333.
- PAYNE, T. H., DETMER, D. E., WYATT, J. C. & BUCHAN, I. E. 2010. National-scale clinical information exchange in the United Kingdom: lessons for the United States. *Journal of the American Medical Informatics Association*, 18, 91-98.
- PEACOCK, J. & PEACOCK, P. 2011. *Oxford handbook of medical statistics*, Oxford University Press.

- PEROTTE, A., RANGANATH, R., HIRSCH, J. S., BLEI, D. & ELHADAD, N. 2015. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22, 872-880.
- PETERSEN, I., JOHNSON, A., ISLAM, A., DUCKWORTH, G., LIVERMORE, D. & HAYWARD, A. 2007a. Protective effect of antibiotics against serious complications of common respiratory tract infections: retrospective cohort study with the UK General Practice Research Database. *Bmj*, 335, 982.
- PETERSEN, I., JOHNSON, A. M., ISLAM, A., DUCKWORTH, G., LIVERMORE, D. M. & HAYWARD, A. C. 2007b. Protective effect of antibiotics against serious complications of common respiratory tract infections: retrospective cohort study with the UK General Practice Research Database. *BMJ*, 335, 982.
- PFUNTNER, A., WIER, L. & STOCKS, C. 2006. Most frequent conditions in US hospitals, 2011: statistical brief# 162.
- PHE, P. H. E. 2013. Historical vaccine development and introduction of routine vaccine programmes in the UK.
- PHE, P. H. E. 2014. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) Report 2014.
- PHE, P. H. E. 2015. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) report (2010 TO 2014).
- PHE, P. H. E. 2016. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) Report 2016.
- PHE, P. H. E. 2017a. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report 2017.
- PHE, P. H. E. 2017b. Management and treatment of
common infections.
- PHE, P. H. E. 2019. Surveillance of surgical site infections in NHS hospitals in England. Public Health England.
- PHUNG, D. T., WANG, Z., RUTHERFORD, S., HUANG, C. & CHU, C. 2013. Body mass index and risk of pneumonia: A systematic review and meta-analysis. *Obesity Reviews*, 14, 839-857.
- PICK, H., DANIEL, P., RODRIGO, C., BEWICK, T., ASHTON, D., LAWRENCE, H., BASKARAN, V., EDWARDS-PRITCHARD, R. C., SHEPPARD, C., ELETU, S. D., ROSE, S., LITT, D., FRY, N. K., LADHANI, S., CHAND, M., TROTTER,

- C., MCKEEVER, T. M. & LIM, W. S. 2020. Pneumococcal serotype trends, surveillance and risk factors in UK adult pneumonia, 2013-18. *Thorax*, 75, 38-49.
- PICK, H., LACEY, J., HODGSON, D., MACDONALD, E., TURVEY, A. & BEWICK, T. 2014. Clinical characteristics of hospitalised patients misdiagnosed with community-acquired pneumonia. *Thorax*, 69, A10.
- PILISHVILI, T. & BENNETT, N. M. 2015. Pneumococcal disease prevention among adults: strategies for the use of pneumococcal vaccines. *Vaccine*, 33, D60-D65.
- PODOLSKY, S. H. 2005. The changing fate of pneumonia as a public health concern in 20th-century America and beyond. *American journal of public health*, 95, 2144-2154.
- PODOLSKY, S. H. 2006. *Pneumonia before antibiotics: therapeutic evolution and evaluation in twentieth-century America*, JHU Press.
- POLVERINO, E., CILLONIZ, C., GABARRUS, A., AMARO, R., DOMINGO, R., SIALER, S., SELLALES, J. & TORRESI, A. 2013. Influence of comorbidities on pneumococcal community-acquired pneumonia. *European Respiratory Journal. Conference: European Respiratory Society Annual Congress*, 42.
- POSTERNAK, M. A. & MILLER, I. 2001. Untreated short-term course of major depression: a meta-analysis of outcomes from studies using wait-list control groups. *Journal of affective disorders*, 66, 139-146.
- POUWELS, K. B., DOLK, F. C. K., SMITH, D. R. M., SMIESZEK, T. & ROBOTHAM, J. V. 2018. Explaining variation in antibiotic prescribing between general practices in the UK. *Journal of Antimicrobial Chemotherapy*, 73, ii27-ii35.
- PRIMARY CARE DOMAIN, N. D. & SERVICE, N. H. 2018. *Appointments in General Practice-October 2018: Summary* [Online]. NHS Digital, part of the Government Statistical Service. Available: <https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/oct-2018> [Accessed 11th June 2019].
- PRINA, E., RANZANI, O. T. & TORRES, A. 2015. Community-acquired pneumonia. *The Lancet*, 386, 1097-1108.
- PRINCIPI, N. & ESPOSITO, S. 2012. Macrolide-resistant *Mycoplasma pneumoniae*: its role in respiratory infection. *Journal of antimicrobial chemotherapy*, 68, 506-511.
- PRYOR, T. A. & HRIPCSAK, G. Sharing MLM's: an experiment between Columbia-Presbyterian and LDS Hospital. Proceedings of the Annual Symposium

on Computer Application in Medical Care, 1993. American Medical Informatics Association, 399.

QUAN, T. P., FAWCETT, N. J., WRIGHTSON, J. M., FINNEY, J., WYLLIE, D., JEFFERY, K., JONES, N., SHINE, B., CLARKE, L. & CROOK, D. 2016a. Increasing burden of community-acquired pneumonia leading to hospitalisation, 1998–2014. *Thorax*, thoraxjnl-2015-207688.

QUAN, T. P., FAWCETT, N. J., WRIGHTSON, J. M., FINNEY, J., WYLLIE, D., JEFFERY, K., JONES, N., SHINE, B., CLARKE, L., CROOK, D., WALKER, A. S., PETO, T. E. & INFECTIONS IN OXFORDSHIRE RESEARCH, D. 2016b. Increasing burden of community-acquired pneumonia leading to hospitalisation, 1998-2014. *Thorax*, 71, 535-42.

QUAN, T. P., FAWCETT, N. J., WRIGHTSON, J. M., FINNEY, J., WYLLIE, D., JEFFERY, K., JONES, N., SHINE, B., CLARKE, L., CROOK, D., WALKER, A. S. & PETO, T. E. A. 2016c. Increasing burden of community-acquired pneumonia leading to hospitalisation, 1998–2014. *Thorax*.

R CORE TEAM 2020. R: A Language and Environment for Statistical Computing. 1.2.5042 ed.

RAJKOMAR, A., OREN, E., CHEN, K., DAI, A. M., HAJAJ, N., HARDT, M., LIU, P. J., LIU, X., MARCUS, J. & SUN, M. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1, 18.

RAVINDRARAJAH, R., HAZRA, N. C., CHARLTON, J., JACKSON, S. H., DREGAN, A. & GULLIFORD, M. C. 2018. Incidence and mortality of fractures by frailty level over 80 years of age: cohort study using UK electronic health records. *BMJ open*, 8, e018836.

RAVINDRARAJAH, R., HAZRA, N. C., HAMADA, S., CHARLTON, J., JACKSON, S. H., DREGAN, A. & GULLIFORD, M. C. 2017. Systolic blood pressure trajectory, frailty, and all-cause mortality > 80 years of age: cohort study using electronic health records. *Circulation*, 135, 2357-2368.

REARDON, S. 2014. WHO warns against 'post-antibiotic' era. *Nature news*.

REHMAN, M. H. & BATOOL, A. 2015. The concept of pattern based data sharing in big data environments. *International Journal of Database Theory and Application*, 8, 11-18.

RILEY, R. D., HAYDEN, J. A., STEYERBERG, E. W., MOONS, K. G., ABRAMS, K., KYZAS, P. A., MALATS, N., BRIGGS, A., SCHROTER, S. & ALTMAN, D. G. 2013. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS medicine*, 10, e1001380.

- RILEY, R. D., MOONS, K. G. M., SNELL, K. I. E., ENSOR, J., HOOFT, L., ALTMAN, D. G., HAYDEN, J., COLLINS, G. S. & DEBRAY, T. P. A. 2019a. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ*, 364, k4597.
- RILEY, R. D., SNELL, K. I., ENSOR, J., BURKE, D. L., HARRELL JR, F. E., MOONS, K. G. & COLLINS, G. S. 2019b. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*, 38, 1276-1296.
- RILEY, R. D., VAN DER WINDT, D., CROFT, P. & MOONS, K. G. 2019c. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*, Oxford University Press.
- RILLERA, D. J. O. & MANUEL, I. 2019. Association of procalcitonin levels and risk of mortality in adults with hospital acquired pneumonia and high risk community acquired pneumonia. *American Journal of Respiratory and Critical Care Medicine. Conference*, 199.
- RIMMER, A. 2014. Doctors must avoid jargon when talking to patients, royal college says. *BMJ*, 348, g4131.
- RIVETT, G. 2014. *National Health Service History* [Online]. Available: http://www.nhshistory.net/a_guide_to_the_nhs.htm [Accessed 21/11/2019 2019].
- ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C. & MÜLLER, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*.
- ROCKWOOD, K. & MITNITSKI, A. 2011. Frailty defined by deficit accumulation and geriatric medicine defined by frailty. *Clinics in geriatric medicine*, 27, 17-26.
- RODGERS, S., AVERY, A. J., MEECHAN, D., BRIANT, S., GERAGHTY, M., DORAN, K. & WHYNES, D. K. 1999. Controlled trial of pharmacist intervention in general practice: the effect on prescribing costs. *Br J Gen Pract*, 49, 717-720.
- RODRIGUEZ-MAÑAS, L. & FRIED, L. P. 2015. Frailty in the clinical scenario. *The Lancet*, 385, e7-e9.
- RÖNNEGÅRD, L., SHEN, X. & ALAM, M. 2010. hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2, 20-28.
- ROSS, M., WEI, W. & OHNO-MACHADO, L. 2014. "Big data" and the electronic health record. *Yearbook of medical informatics*, 23, 97-104.

ROSSI, G. A., MEDICI, M. C., ARCANGELETTI, M. C., LANARI, M., MEROLLA, R., PAPANATTI, U. D. L., SILVESTRI, M., PISTORIO, A., CHEZZI, C. & GROUP, O. R. S. 2007. Risk factors for severe RSV-induced lower respiratory tract infection over four consecutive epidemics. *European journal of pediatrics*, 166, 1267-1272.

ROTMAN, D. 2020. *We're not prepared for the end of Moore's Law* [Online]. MIT Technology Review. Available: <https://www.technologyreview.com/2020/02/24/905789/were-not-prepared-for-the-end-of-moores-law/> [Accessed 29/07 2020].

ROYSTON, P., ALTMAN, D. G. & SAUERBREI, W. 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25, 127-141.

ROYSTON, P., MOONS, K. G., ALTMAN, D. G. & VERGOUWE, Y. 2009. Prognosis and prognostic research: developing a prognostic model. *Bmj*, 338, b604.

ROYSTON, P. & SAUERBREI, W. 2008. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*, John Wiley & Sons.

RUIZ, J., PONS, M. J. & GOMES, C. 2012. Transferable mechanisms of quinolone resistance. *International journal of antimicrobial agents*, 40, 196-203.

RUUSKANEN, O., LAHTI, E., JENNINGS, L. C. & MURDOCH, D. R. 2011. Viral pneumonia. *The Lancet*, 377, 1264-1275.

SADIGOV, A. & ABDULLAYEV, F. 2017. Correlation between emphysema and pneumonia risk in patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS*, 195.

SAHUQUILLO-ARCE, J. M., MENÉNDEZ, R., MÉNDEZ, R., AMARA-ELORI, I., ZALACAIN, R., CAPELASTEGUI, A., ASPA, J., BORDERÍAS, L., MARTÍN-VILLASCLARAS, J. J. & BELLO, S. 2016. Age-related risk factors for bacterial aetiology in community-acquired pneumonia. *Respirology*, 21, 1472-1479.

SALLUH, J. I., SHINOTSUKA, C. R., SOARES, M., BOZZA, F. A., LAPA E SILVA, J. R., TURA, B. R., BOZZA, P. T. & VIDAL, C. G. 2010. Cortisol levels and adrenal response in severe community-acquired pneumonia: a systematic review of the literature. *Journal of Critical Care*, 25, 541.e1-8.

SAMOKHVALOV, A. V., IRVING, H. M. & REHM, J. 2010. Alcohol consumption as a risk factor for pneumonia: A systematic review and meta-analysis. *Epidemiology and Infection*, 138, 1789-1795.

SANZ, F., RESTREPO, M. I., FERNANDEZ-FABRELLAS, E., CERVERA, A., BRIONES, M. L., NOVELLA, L., AGUAR, M. C., CHINER, E., FERNANDEZ, J. F. & BLANQUER, J. 2014a. Does prolonged onset of symptoms have a prognostic significance in community-acquired pneumonia? *Respirology*.

SANZ, F., RESTREPO, M. I., FERNÁNDEZ-FABRELLAS, E., CERVERA, Á., BRIONES, M. L., NOVELLA, L., AGUAR, M. C., CHINER, E., FERNANDEZ, J. F. & BLANQUER, J. 2014b. Does prolonged onset of symptoms have a prognostic significance in community-acquired pneumonia? *Respirology*, 19, 1073-1079.

SANZ HERRERO, F. & BLANQUER OLIVAS, J. 2012. Microbiology and risk factors for community-acquired pneumonia. *Seminars in Respiratory & Critical Care Medicine*, 33, 220-31.

SCAMBLER, G. 2008. *Sociology as Applied to Medicine E-Book*, Elsevier Health Sciences.

SCHAFFER, J. 2015. What not to multiply without necessity. *Australasian Journal of Philosophy*, 93, 644-664.

SCHAPPERT, S. M. & BURT, C. W. 2006. Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 2001-02. *Vital and Health Statistics. Series 13, Data from the National Health Survey*, 1-66.

SCHEMBRI, S., WILLIAMSON, P. A., SHORT, P. M., SINGANAYAGAM, A., AKRAM, A., TAYLOR, J., SINGANAYAGAM, A., HILL, A. T. & CHALMERS, J. D. 2013. Cardiovascular events after clarithromycin use in lower respiratory tract infections: analysis of two prospective cohort studies. *Bmj*, 346, f1235.

SCHEPP, S. K., TIRSCHWELL, D. L. & LONGSTRETH, W. T. 2012. A clinical prediction rule for pneumonia after acute stroke. *Stroke. Conference*, 43.

SCHIERENBERG, A., MINNAARD, M. C., HOPSTAKEN, R. M., VAN DE POL, A. C., BROEKHUIZEN, B. D. L., DE WIT, N. J., REITSMA, J. B., VAN VUGT, S. F., GRAFFELMAN, A. W., MELBYE, H., RAINER, T. H., STEURER, J., HOLM, A., GONZALES, R., DINANT, G. J., DE GROOT, J. A. H. & VERHEIJ, T. J. M. 2016. External validation of prediction models for pneumonia in primary care patients with lower respiratory tract infection: An individual patient data meta-analysis. *PLoS ONE*, 11.

SCHLOEFFEL, P. 2004. ISO/DTR 20514. *Health informatics-electronic health record: definition, scope and context. Fourth Draft*.

SCHROEDER, S. A. 2004. Tobacco Control in the Wake of the 1998 Master Settlement Agreement. *New England Journal of Medicine*, 350, 293-301.

- SCHUBERT, K., JASPER, E., MANDRAKA, F., GEPPERT, R. & REUTER, S. 2013. Observational study on diagnostics and treatment in community acquired pneumonia (CAP). *International Journal of Medical Microbiology*, 303, 87.
- SCHUETZ, P., SUTER-WIDMER, I., CHAUDRI, A., CHRIST-CRAIN, M., ZIMMERLI, W. & MUELLER, B. 2011. Prognostic value of procalcitonin in community-acquired pneumonia. *European Respiratory Journal*, 37, 384-392.
- SCHWABER, M. J., LEV, B., ISRAELI, A., SOLTER, E., SMOLLAN, G., RUBINOVITCH, B., SHALIT, I. & CARMELI, Y. 2011. Containment of a country-wide outbreak of carbapenem-resistant *Klebsiella pneumoniae* in Israeli hospitals via a nationally implemented intervention. *Clinical infectious diseases*, cir025.
- SEVENSTER, M., BUURMAN, J., LIU, P., PETERS, J. & CHANG, P. 2015. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. *Applied clinical informatics*, 6, 600-610.
- SHAH, N. D., STEYERBERG, E. W. & KENT, D. M. 2018. Big data and predictive analytics: recalibrating expectations. *Jama*, 320, 27-28.
- SHAH, N. H., LEPENDU, P., BAUER-MEHREN, A., GHEBREMARIAM, Y. T., IYER, S. V., MARCUS, J., NEAD, K. T., COOKE, J. P. & LEEPER, N. J. 2015. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *PloS one*, 10, e0124653.
- SHAMEER, K., GLICKSBERG, B. S., HODOS, R., JOHNSON, K. W., BADGELEY, M. A., READHEAD, B., TOMLINSON, M. S., O'CONNOR, T., MIOTTO, R. & KIDD, B. A. 2017. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Briefings in bioinformatics*, 19, 656-678.
- SHAMEER, K., JOHNSON, K. W., GLICKSBERG, B. S., DUDLEY, J. T. & SENGUPTA, P. P. 2018. Machine learning in cardiovascular medicine: are we there yet? *Heart*, 104, 1156-1164.
- SHEEHAN, K. J., WILLIAMSON, L., ALEXANDER, J., FILLITER, C., SOBOLEV, B., GUY, P., BEARNE, L. M. & SACKLEY, C. 2018. Prognostic factors of functional outcome after hip fracture surgery: a systematic review. *Age and Ageing*, 47, 661-670.
- SHIVANI, P. & CPRD. 2017. *CPRD GOLD Data Specification (Version 2.0)* [Online]. Available: https://cprd.com/sites/default/files/CPRD%20GOLD%20Full%20Data%20Specification%20v2.0_0.pdf [Accessed 17/12/2019].

SIBISAKKARAVARTHI, M. & SUBRAMANIAM, P. 2017. Artificial Intelligence Network Load Comparison in Ant Colony Optimization.

SIDDAWAY, A. P., WOOD, A. M. & HEDGES, L. V. 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, 70, 747-770.

SILJAN, W. W., HOLTER, J. C., MICHELSEN, A. E., NYMO, S. H., NORSETH, J., HUSEBYE, E., LAURITZEN, T., UELAND, T., MOLLNES, T. E., AUKRUST, P. & HEGGELUND, L. 2018. Procalcitonin has wider applicability in community-acquired pneumonia than other biomarkers. *European Respiratory Journal. Conference: European Respiratory Society International Congress, ERS*, 52.

SIMONETTI, A. F., VIASUS, D., GARCIA-VIDAL, C. & CARRATALA, J. 2014a. Management of community-acquired pneumonia in older adults. *Therapeutic Advances in Infectious Disease*, 2, 3-16.

SIMONETTI, A. F., VIASUS, D., GARCIA-VIDAL, C., GRILLO, S., MOLERO, L., DORCA, J. & CARRATALA, J. 2014b. Impact of pre-hospital antibiotic use on community-acquired pneumonia. *Clinical Microbiology and Infection*, 20, O531-O537.

SING, T., SANDER, O., BEERENWINKEL, N. & LENGAUER, T. 2005. ROCr: visualizing classifier performance in R.

SMITH, C. J., BRAY, B. D., HOFFMAN, A., MEISEL, A., HEUSCHMANN, P. U., WOLFE, C. D., RUDD, A. G. & TYRRELL, P. J. 2014. Infections diseases and stroke a novel point-of-care clinical risk score for predicting pneumonia in acute stroke care: a uk multicenter cohort study. *International journal of stroke.*, 9, 215.

SMITH, D. R. M., DOLK, F. C. K., SMIESZEK, T., ROBOTHAM, J. V. & POUWELS, K. B. 2018. Understanding the gender gap in antibiotic prescribing: a cross-sectional analysis of English primary care. *BMJ Open*, 8.

SO, A. D. & WOODHOUSE, W. 2010. Thailand's Antibiotic Smart Use Initiative.

SPINDLER, C., STRALIN, K., ERIKSSON, L., HJERDT-GOSCINSKI, G., HOLMBERG, H., LIDMAN, C., NILSSON, A., ORTQVIST, A., HEDLUND, J. & COMMUNITY ACQUIRED PNEUMONIA WORKING GROUP OF THE SWEDISH SOCIETY OF INFECTIOUS, D. 2012. Swedish guidelines on the management of community-acquired pneumonia in immunocompetent adults-- Swedish Society of Infectious Diseases 2012. *Scandinavian Journal of Infectious Diseases*, 44, 885-902.

- SRIVASTAVA, S. A. 2010. Seasonal predictive factors of acute respiratory tract infections in children. *Thorax*, 65, 106-106.
- STATISTICAL RESEARCH AND APPLICATIONS BRANCH, N. C. I. 2018. Joinpoint Regression Program. Version 4.6.0.0 ed.
- STEELE, A. J., DENAXAS, S. C., SHAH, A. D., HEMINGWAY, H. & LUSCOMBE, N. M. 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*, 13, e0202344.
- STERN, A., SKALSKY, K., AVNI, T., CARRARA, E., LEIBOVICI, L. & PAUL, M. 2017. Corticosteroids for pneumonia. *Cochrane Database of Systematic Reviews*, 2017.
- STERNE, J. A., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009a. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, b2393.
- STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009b. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, b2393.
- STEYERBERG, E. W. 2008. *Clinical prediction models: a practical approach to development, validation, and updating*, Springer Science & Business Media.
- STEYERBERG, E. W., MOONS, K. G., VAN DER WINDT, D. A., HAYDEN, J. A., PEREL, P., SCHROTER, S., RILEY, R. D., HEMINGWAY, H., ALTMAN, D. G. & GROUP, P. 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS medicine*, 10, e1001381.
- STEYERBERG, E. W. & VERGOUWE, Y. 2014. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal*, 35, 1925-1931.
- STEYERBERG, E. W., VICKERS, A. J., COOK, N. R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M. J. & KATTAN, M. W. 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21, 128.
- STOCKS, N. & FAHEY, T. 2002. Labelling of acute respiratory illness: evidence of between-practitioner variation in the UK. *Family Practice*, 19, 375-377.

- STOLZ, D. 2016. Procalcitonin in Severe Community-Acquired Pneumonia: Some Precision Medicine Ready for Prime Time. *Chest*, 150, 769-771.
- STONE, S. P., FULLER, C., SAVAGE, J., COOKSON, B., HAYWARD, A., COOPER, B., DUCKWORTH, G., MICHIE, S., MURRAY, M. & JEANES, A. 2012. Evaluation of the national Cleanyourhands campaign to reduce *Staphylococcus aureus* bacteraemia and *Clostridium difficile* infection in hospitals in England and Wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study.
- STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. & ZEILEIS, A. 2008. Conditional variable importance for random forests. *BMC bioinformatics*, 9, 307.
- STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. & HOTHORN, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8, 25.
- SUAYA, J. A., JIANG, Q., SCOTT, D. A., GRUBER, W. C., WEBBER, C., SCHMOELE-THOMA, B., HALL-MURRAY, C. K., JODAR, L. & ISTURIZ, R. E. 2018. Post hoc analysis of the efficacy of the 13-valent pneumococcal conjugate vaccine against vaccine-type community-acquired pneumonia in at-risk older adults. *Vaccine*, 36, 1477-1483.
- SUBRAMANIAN, D. N., MUSONDA, P., SANKARAN, P., TARIQ, S. M., KAMATH, A. V. & MYINT, P. K. 2013. Performance of SOAR (systolic blood pressure, oxygenation, age and respiratory rate) scoring criteria in community-acquired pneumonia: A prospective multi-centre study. *Age and Ageing*, 42, 94-97.
- SULLIVAN, P. & GOLDMANN, D. 2011. The promise of comparative effectiveness research. *Jama*, 305, 400-401.
- SUN, X., DOUIRI, A. & GULLIFORD, M. 2019. Pneumonia incidence trends in UK primary care from 2002 to 2017: population-based cohort study. *Epidemiology & Infection*, 147.
- SUN, X. & GULLIFORD, M. C. 2019. Reducing antibiotic prescribing in primary care in England from 2014 to 2017: population-based cohort study. *BMJ open*, 9, e023989.
- SUNGUR BITEKER, F., CIL, C., CELIK, O., OZLEK, B., OZLEK, E. & GOKCEK, A. 2019. Right Heart Function in Community-Acquired Pneumonia. *Heart Lung and Circulation*, 28, e145.

- TAN, T., LITTLE, P., STOKES, T. & GUIDELINE DEVELOPMENT, G. 2008. Antibiotic prescribing for self limiting respiratory tract infections in primary care: summary of NICE guidance. *BMJ*, 337, a437.
- TANDAY, S. 2013. C-reactive protein could predict pneumonia in COPD. *The Lancet*, Respiratory medicine. 1, 510.
- TANG, P. C., ASH, J. S., BATES, D. W., OVERHAGE, J. M. & SANDS, D. Z. 2006. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association*, 13, 121-126.
- TANZELLA, G., MOTOS, A., BATTAGLINI, D., MELI, A. & TORRES, A. 2019. Optimal approaches to preventing severe community-acquired pneumonia. *Expert Review of Respiratory Medicine*, 13, 1005-1018.
- TAO, C., PARKER, C. G., ONIKI, T. A., PATHAK, J., HUFF, S. M. & CHUTE, C. G. An OWL meta-ontology for representing the clinical element model. AMIA annual symposium proceedings, 2011. American Medical Informatics Association, 1372.
- TASHIRO, H., KIKUTANI, T., TAMURA, F., TAKAHASHI, N., TOHARA, T., NAWACHI, K., MAEKAWA, K. & KUBOKI, T. 2019. Relationship between oral environment and development of pneumonia and acute viral respiratory infection in dependent older individuals. *Geriatrics and Gerontology International*, 19, 1136-1140.
- TATE, A. R., WILLIAMS, T., PURI, S., BELOFF, N. & VAN STAA, T. Developing quality scores for electronic health records for clinical research: a study using the General Practice Research Database. Proceedings of the first international workshop on Managing interoperability and complexity in health systems, 2011. ACM, 35-42.
- TEAM, R. C. 2013. R: A language and environment for statistical computing.
- THANGARATINAM, S. & REDMAN, C. W. 2005. The delphi technique. *The obstetrician & gynaecologist*, 7, 120-125.
- THE MATHWORKS, I. 2020. Classification Learner.
- THERNEAU, T. & ATKINSON, B. 2019. rpart: Recursive Partitioning and Regression Trees.
- THORNTON SNIDER, J., LUNA, Y., WONG, K. S., ZHANG, J., CHEN, S. S., GLESS, P. J. & GOLDMAN, D. P. 2012. Inhaled corticosteroids and the risk of

pneumonia in Medicare patients with COPD. *Current Medical Research and Opinion*, 28, 1959-1967.

TICINESI, A., NOUVENNE, A., FOLESANI, G., PRATI, B., MORELLI, I., GUIDA, L., LAURETANI, F., MAGGIO, M. & MESCHI, T. 2016. An investigation of multimorbidity measures as risk factors for pneumonia in elderly frail patients admitted to hospital. *European Journal of Internal Medicine*, 28, 102-106.

TORRES, A., BLASI, F., PEETERMANS, W., VIEGI, G. & WELTE, T. 2014. The aetiology and antibiotic management of community-acquired pneumonia in adults in Europe: a literature review. *European journal of clinical microbiology & infectious diseases*, 33, 1065-1079.

TORRES, A., PEETERMANS, W. E., VIEGI, G. & BLASI, F. 2013. Risk factors for community-acquired pneumonia in adults in Europe: A literature review. *Thorax*, 68, 1057-1065.

TRICCO, A. C., LILLIE, E., ZARIN, W., O'BRIEN, K. K., COLQUHOUN, H., LEVAC, D., MOHER, D., PETERS, M. D., HORSLEY, T. & WEEKS, L. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine*, 169, 467-473.

TRIFIRO, G. 2011. Antipsychotic drug use and community-acquired pneumonia. *Current Infectious Disease Reports*, 13, 262-268.

TROEGER, C., BLACKER, B., KHALIL, I. A., RAO, P. C., CAO, J., ZIMSEN, S. R., ALBERTSON, S. B., DESHPANDE, A., FARAG, T. & ABEBE, Z. 2018. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Infectious Diseases*.

TROEGER, C., FOROUZANFAR, M., RAO, P. C., KHALIL, I., BROWN, A., SWARTZ, S., FULLMAN, N., MOSSER, J., THOMPSON, R. L. & REINER JR, R. C. 2017. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17, 1133-1161.

TROTTER, C. L., STUART, J. M., GEORGE, R. & MILLER, E. 2008. Increasing hospital admissions for pneumonia, England. *Emerging infectious diseases*, 14, 727.

UDDIN, M. F. & GUPTA, N. Seven V's of Big Data understanding Big Data to extract value. Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education, 2014. IEEE, 1-5.

VAMOS, E. P., PAPE, U. J., BOTTLE, A., HAMILTON, F. L., CURCIN, V., NG, A., MOLOKHIA, M., CAR, J., MAJEED, A. & MILLETT, C. 2011. Association of practice size and pay-for-performance incentives with the quality of diabetes management in primary care. *CMAJ*, 183, E809-E816.

VAN BUYNDER, P. 2019. Reducing pneumococcal risk in people aged 65 years and over. *Medicine Today*, 20, 11-14.

VAN DEN BROEK D'OBRENAN, J., VERHEIJ, T. J., NUMANS, M. E. & VAN DER VELDEN, A. W. 2014. Antibiotic use in Dutch primary care: relation between diagnosis, consultation and treatment. *Journal of Antimicrobial Chemotherapy*, 69, 1701-1707.

VAN DRIEST, S. L., WELLS, Q. S., STALLINGS, S., BUSH, W. S., GORDON, A., NICKERSON, D. A., KIM, J. H., CROSSLIN, D. R., JARVIK, G. P. & CARRELL, D. S. 2016. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *Jama*, 315, 47-57.

VAN STAA, T.-P., WEGMAN, S., DE VRIES, F., LEUFKENS, B. & COOPER, C. 2001. Use of statins and risk of fractures. *Jama*, 285, 1850-1855.

VAN VUGT, S. F., VERHEIJ, T. J. M., DE JONG, P. A., BUTLER, C. C., HOOD, K., COENEN, S., GOOSSENS, H., LITTLE, P. & BROEKHUIZEN, B. D. L. 2013. Diagnosing pneumonia in patients with acute cough: Clinical judgment compared to chest radiography. *European Respiratory Journal*, 42, 1076-1082.

VARIAN, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28, 3-28.

VIASUS, D., GARCIA-VIDAL, C., CASTELLOTE, J., ADAMUZ, J., VERDAGUER, R., DORCA, J., MANRESA, F., GUDIOL, F. & CARRATALA, J. 2011. Community-acquired pneumonia in patients with liver cirrhosis: Clinical features, outcomes, and usefulness of severity scores. *Medicine*, 90, 110-118.

VIASUS, D., GARCIA-VIDAL, C., CASTELLOTE, J., ADAMUZ, J., VERDAGUER, R., GUDIOL, F. & CARRATALA, J. 2010. Epidemiology, clinical features, and outcomes of community-acquired pneumonia in patients with liver cirrhosis. *Clinical Microbiology and Infection*, 16, S381.

VICKERS, A. J. & ELKIN, E. B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26, 565-574.

VICKERS, A. J., VAN CALSTER, B. & STEYERBERG, E. W. 2016. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, i6.

- VILA-CORCOLES, A., OCHOA-GONDAR, O., RODRIGUEZ-BLANCO, T., RAGA-LURIA, X., GOMEZ-BERTOMEU, F. & GROUP, E. S. 2009. Epidemiology of community-acquired pneumonia in older adults: a population-based study. *Respiratory medicine*, 103, 309-316.
- VILANOVA, M. B., FALGUERA, M., PENA, M., SANCHEZ, V., CHICA, I., MONTSERRAT-CAPDEVILA, J., ESQUINAS, C. & MARSAL, J. R. 2012. Obesity and metabolic syndrome as risk factors for community-acquired pneumonia. *Clinical Microbiology and Infection*, 18, 136-137.
- VINOGRADOVA, Y., HIPPISEY-COX, J. & COUPLAND, C. 2009. Identification of new risk factors for pneumonia: population-based case-control study. *Br J Gen Pract*, 59, e329-e338.
- VIOLI, F., CANGEMI, R., FALCONE, M., TALIANI, G., PIERALLI, F., VANNUCCHI, V., NOZZOLI, C., VENDITTI, M., CHIRINOS, J. A. & CORRALES-MEDINA, V. F. 2017. Cardiovascular complications and short-term mortality risk in community-acquired pneumonia. *Clinical Infectious Diseases*, 64, 1486-1493.
- VITTINGHOFF, E. & MCCULLOCH, C. E. 2006. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, 165, 710-718.
- WAGHOLIKAR, K. B., SUNDARARAJAN, V. & DESHPANDE, A. W. 2012. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, 36, 3029-3049.
- WALDROP, M. M. 2016. The chips are down for Moore's law. *Nature News*, 530, 144.
- WALTERS, G., SUM LIU, H., GEMZA, M. & RAUF, F. 2011. Does the serum C-reactive protein (CRP) predict adverse outcomes in patients admitted with community acquired pneumonia? *European Respiratory Journal. Conference: European Respiratory Society Annual Congress*, 38.
- WARDLAW, T., SALAMA, P., JOHANSSON, E. W. & MASON, E. 2006. Pneumonia: the leading killer of children. *The Lancet*, 368, 1048-1050.
- WATANABE TEJADA, L. C., PAJE, D., SHAKEEL, Q. L., UDUMAN, A. K., VAHIA, A. & CABRERA, R. 2013. Effect of comorbidities on clinical outcomes in low-risk curb-65 patients. *Journal of General Internal Medicine*, 28, S67.
- WATKINS, R. R. & LEMONOVICH, T. L. 2011. Diagnosis and management of community-acquired pneumonia in adults. *American family physician*, 83, 1299-1306.

WATSON, D. S., KRUTZINNA, J., BRUCE, I. N., GRIFFITHS, C. E., MCINNES, I. B., BARNES, M. R. & FLORIDI, L. 2019. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, 364, 1886.

WEINGARTEN, S. R., LLOYD, L., CHIOU, C.-F. & BRAUNSTEIN, G. D. 2002. Do subspecialists working outside of their specialty provide less efficient and lower-quality care to hospitalized patients than do primary care physicians? *Archives of Internal Medicine*, 162, 527-532.

WEISS, A. J., WIER, L. M., STOCKS, C. & BLANCHARD, J. 2006. Overview of emergency department visits in the United States, 2011: statistical brief# 174.

WELTE, T., TORRES, A. & NATHWANI, D. 2012. Clinical and economic burden of community-acquired pneumonia among adults in Europe. *Thorax*, 67, 71-79.

WENG, S. F., REPS, J., KAI, J., GARIBALDI, J. M. & QURESHI, N. 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12, e0174944.

WENG, S. F., VAZ, L., QURESHI, N. & KAI, J. 2019. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PloS one*, 14, e0214365.

WEST, J. B. 2012. *Respiratory physiology: the essentials*, Lippincott Williams & Wilkins.

WHO 2005. *Good clinical diagnostic practice: a guide for clinicians in developing countries to the clinical diagnosis of disease and to making proper use of clinical diagnostic services*, World Health Organization.

WHO. 2016a. *Human Factors: Technical Series on Safer Primary Care* [Online]. World Health Organisation Available: <https://apps.who.int/iris/bitstream/handle/10665/252273/9789241511612-eng.pdf;jsessionid=6724DA99A8EB21E8960681F46B47CB6A?sequence=1> [Accessed 11/09/2020].

WHO 2016b. Pneumonia.

WHO. 2018a. *Protocol for surgical site infection surveillance with a focus on settings with limited resources* [Online]. Available: <https://www.who.int/infection-prevention/tools/surgical/SSI-surveillance-protocol.pdf> [Accessed 6th June 2019].

WHO. 2011. *World Health Day 2011*

Combat drug resistance: no action today means no cure tomorrow [Online]. Available:

http://www.who.int/mediacentre/news/statements/2011/whd_20110407/en/
[Accessed 10st August 2016].

WHO. 2013. The evolving threat of antimicrobial resistance: options for action. 2012. *Geneva, Switzerland: World Health Organization Google Scholar*.

WHO. 2018b. World Pneumonia Day.

WHO. 2019a. *Classifications-ICD* [Online]. Available:
<https://www.who.int/classifications/en/> [Accessed 10/01/2020].

WHO. 2019b. *Influenza: vaccines* [Online]. Available:
<https://www.who.int/influenza/vaccines/about/en/> [Accessed 16/10/2019 2019].

WHO. 2020. *International Classification of Primary Care, Second edition (ICPC-2)* [Online]. Available: <https://www.who.int/classifications/icd/adaptations/icpc2/en/>
[Accessed 10/01/2020].

WICKHAM, H. 2016a. *ggplot2: elegant graphics for data analysis*, Springer.

WICKHAM, H. 2016b. *ggplot2: Elegant Graphics for Data Analysis*.

WIERSINGA, W. J., BONTEN, M. J., BOERSMA, W. G., JONKERS, R. E., ALEVA, R. M., KULLBERG, B. J., SCHOUTEN, J. A., DEGENER, J. E., JANKNEGT, R., VERHEIJ, T. J., SACHS, A. P. E. & PRINS, J. M. 2012. SWAB/NVALT (dutch working party on antibiotic policy and dutch association of chest physicians) guidelines on the management of community-acquired pneumonia in adults. *Netherlands Journal of Medicine*, 70, 90-101.

WILLIAMS, N., COOMBS, N. A., RIGGE, L., JOSEPHS, L., JOHNSON, M., THOMAS, D. M. & WILKINSON, T. M. A. 2015. Co-morbidity and pneumonia risk in COPD patients: A population database analysis of primary care patients. *Thorax*, 70, A69.

WILLIAMS, S., HALLS, A., TONKIN-CRINE, S., MOORE, M., LATTER, S., LITTLE, P., EYLES, C., POSTLE, K. & LEYDON, G. 2017. General practitioner and nurse prescriber experiences of prescribing antibiotics for respiratory tract infections in UK primary care out-of-hours services (the UNITE study). *Journal of Antimicrobial Chemotherapy*.

WILLIAMS, T., VAN STAA, T., PURI, S. & EATON, S. 2012. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic advances in drug safety*, 3, 89-99.

WIRTH, N. 1986. Algorithms and data structures.

WOLF, A. 2018. *RE: CPRD00023564 Free text access*. Type to SUN, X.

- WOLF, A., DEDMAN, D., CAMPBELL, J., BOOTH, H., LUNN, D., CHAPMAN, J. & MYLES, P. 2019. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International journal of epidemiology*.
- WOLPERT, D. H. & MACREADY, W. G. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1, 67-82.
- WONCA, W. O. O. F. D. 2016. International Classification of Primary Care.
- WOODHEAD, M., BLASI, F., EWIG, S., GARAU, J., HUCHON, G., IEVEN, M., ORTQVIST, A., SCHABERG, T., TORRES, A. & VAN DER HEIJDEN, G. 2011a. Guidelines for the management of adult lower respiratory tract infections-Full version. *Clinical microbiology and infection*, 17, E1-E59.
- WOODHEAD, M., BLASI, F., EWIG, S., GARAU, J., HUCHON, G., IEVEN, M., ORTQVIST, A., SCHABERG, T., TORRES, A., VAN DER HEIJDEN, G., READ, R. & VERHEIJ, T. J. M. 2011b. Guidelines for the management of adult lower respiratory tract infections - Full version. *Clinical Microbiology and Infection*, 17, E1-E59.
- WORLD HEALTH ORGANIZATION, W. 2018. *International Statistical Classification of Diseases and Related Health Problems 10th Revision 2010* [Online]. Geneva: World Health Organization. Available: <https://icd.who.int/browse10/2010/en> [Accessed 28th November 2018].
- WUNDERINK, R. G. & WATERER, G. 2017. Advances in the causes and management of community acquired pneumonia in adults. *Bmj*, 358, j2471.
- XU, H., ALDRICH, M. C., CHEN, Q., LIU, H., PETERSON, N. B., DAI, Q., LEVY, M., SHAH, A., HAN, X. & RUAN, X. 2014. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*, 22, 179-191.
- XU, H., GASPARINI, A., ISHIGAMI, J., MZAYEN, K., SU, G., BARANY, P., ARNLOV, J., LINDHOLM, B., ELINDER, C. G., MATSUSHITA, K. & CARRERO, J. J. 2017. eGFR and the Risk of Community-Acquired Infections. *Clinical Journal of The American Society of Nephrology: CJASN*, 12, 1399-1408.
- YANBAEVA, D. G., DENTENER, M. A., CREUTZBERG, E. C., WESSELING, G. & WOUTERS, E. F. 2007. Systemic effects of smoking. *Chest*, 131, 1557-1566.
- YAO, L., ZHANG, Y., LI, Y., SANSEAU, P. & AGARWAL, P. 2011. Electronic health records: Implications for drug discovery. *Drug discovery today*, 16, 594-599.

YENDE, S., ALVAREZ, K., LOEHR, L., FOLSOM, A. R., NEWMAN, A. B., WEISSFELD, L. A., WUNDERINK, R. G., KRITCHEVSKY, S. B., MUKAMAL, K. J., LONDON, S. J., HARRIS, T. B., BAUER, D. C., ANGUS, D. C., ATHEROSCLEROSIS RISK IN COMMUNITIES STUDY, T. C. H. S., THE HEALTH, A. & BODY COMPOSITION, S. 2013. Epidemiology and long-term clinical and biologic risk factors for pneumonia in community-dwelling older Americans: analysis of three cohorts. *Chest*, 144, 1008-1017.

YENDE, S., VAN DER POLL, T., LEE, M., HUANG, D. T., NEWMAN, A. B., KONG, L., KELLUM, J. A., HARRIS, T. B., BAUER, D., SATTERFIELD, S., ANGUS, D. C., GENIMS & HEALTH, A. B. C. S. 2010. The influence of pre-existing diabetes mellitus on the host immune response and outcome of pneumonia: analysis of two multicentre cohort studies. *Thorax*, 65, 870-7.

ZHANG, D., YANG, D. & MAKAM, A. N. 2019. Utility of Blood Cultures in Pneumonia. *American Journal of Medicine*, 132, 1233-1238.

ZHANG, J. & KAI, F. Y. 1998. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama*, 280, 1690-1691.

ZHANG, Z. 2017. Survival analysis in the presence of competing risks. *Annals of translational medicine*, 5.

ZHAO, Y. J., LIN, L., ZHOU, H. J., TAN, K. T., CHEW, A. P., FOO, C. G., OH, C. T. D., LIM, B. P. & LIM, W. S. 2016. Cost-effectiveness modelling of novel oral anticoagulants incorporating real-world elderly patients with atrial fibrillation. *International journal of cardiology*, 220, 794-801.

ZHENG, Y.-Y., MA, Y.-T., ZHANG, J.-Y. & XIE, X. 2020. COVID-19 and the cardiovascular system. *Nature Reviews Cardiology*, 17, 259-260.

Appendices:

Appendix A: Comparisons of clinical terminology systems

Table A 1: Comparisons within primary care terminology systems (Read code version 2 and International Classification for Primary Care, ICPC-2)

Read code-2 Chapters		ICPC-2 Chapters
0....	Occupations	A General and unspecified
1....	History / symptoms	A General and unspecified
2....	Examination / Signs	A General and unspecified
3....	Diagnostic procedures	NA
4....	Laboratory procedures	NA
5....	Radiology/physics in medicine	NA
6....	Preventive procedures	NA
7....	Operations, procedures, sites	NA
8....	Other therapeutic procedures	NA
9....	Administration	NA
A....	Infectious and parasitic diseases	NA
B....	Neoplasms	NA
C....	Endocrine, nutritional, metabolic and immunity disorders	T Endocrine, metabolic and nutritional
D....	Diseases of blood and blood-forming organs	B Blood, blood forming organs, lymphatics, spleen
E....	Mental disorders	P Psychological
F....	Nervous system and sense organ diseases	F Eye; H Ear; N Neurological
G....	Circulatory system diseases	K Circulatory
H....	Respiratory system diseases	R Respiratory
J....	Digestive system diseases	D Digestive
K....	Genitourinary system diseases	U Urology; X Female genital system and breast; Y Male genital system
L....	Complications of pregnancy, childbirth and the puerperium	
M....	Skin and subcutaneous tissue diseases	S Skin
N....	Musculoskeletal and connective tissue diseases	L Musculoskeletal
P....	Congenital anomalies	

Q....	Perinatal conditions	W Pregnancy, childbirth, family planning
R....	[D]Symptoms, signs and ill-defined conditions	
S....	Injury and poisoning	NA
T....	Causes of injury and poisoning	NA
U....	[X]External causes of morbidity and mortality	NA
Z....	Unspecified conditions	A General and unspecified
	NA	Z Social problems

Table A 2: Comparison of main structure between Read Code-2 and International Classification of Disease, ICD-10

Read Code-2 Chapters		ICD-10 Chapters
0....	Occupations	NA
1....	History / symptoms	NA
2....	Examination / Signs	NA
3....	Diagnostic procedures	NA
4....	Laboratory procedures	NA
5....	Radiology/physics in medicine	NA
6....	Preventive procedures	NA
7....	Operations, procedures, sites	NA
8....	Other therapeutic procedures	NA
9....	Administration	NA
A....	Infectious and parasitic diseases	I Certain infectious and parasitic diseases
B....	Neoplasms	II Neoplasms
C....	Endocrine, nutritional, metabolic and immunity disorders	IV Endocrine, nutritional and metabolic diseases
D....	Diseases of blood and blood-forming organs	III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E....	Mental disorders	V Mental and behavioural disorders
F....	Nervous system and sense organ diseases	VI Diseases of the nervous system & VII Diseases of the eye and adnexa (F4...) & VIII Diseases of the ear and mastoid process (F5...)
G....	Circulatory system diseases	IX Diseases of the circulatory system
H....	Respiratory system diseases	X Diseases of the respiratory system
J....	Digestive system diseases	XI Diseases of the digestive system

K....	Genitourinary system diseases	XIV Diseases of the genitourinary system
L....	Complications of pregnancy, childbirth and the puerperium	XV Pregnancy, childbirth and the puerperium
M....	Skin and subcutaneous tissue diseases	XII Diseases of the skin and subcutaneous tissue
N....	Musculoskeletal and connective tissue diseases	XIII Diseases of the musculoskeletal system and connective tissue
P....	Congenital anomalies	XVII Congenital malformations, deformations and chromosomal abnormalities
Q....	Perinatal conditions	XVI Certain conditions originating in the perinatal period
R....	[D]Symptoms, signs and ill-defined conditions	XXII Codes for special purposes
S....	Injury and poisoning	XIX Injury, poisoning and certain other consequences of external causes
T....	Causes of injury and poisoning	XX External causes of morbidity and mortality
U....	[X]External causes of morbidity and mortality	XX External causes of morbidity and mortality
Z....	Unspecified conditions	XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
	NA	XXI Factors influencing health status and contact with health services

Table A 3: Comparison of main structure between ICPC-2 and ICD-10

ICPC-2 Chapters		ICD-10 Chapters
	NA	I Certain infectious and parasitic diseases
	NA	II Neoplasms
B	Blood, blood forming organs, lymphatics, spleen	III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
T	Endocrine, metabolic and nutritional	IV Endocrine, nutritional and metabolic diseases
P	Psychological	V Mental and behavioural disorders
N	Neurological	VI Diseases of the nervous system
F	Eye	VII Diseases of the eye and adnexa
H	Ear	VIII Diseases of the ear and mastoid process
K	Circulatory	IX Diseases of the circulatory system
R	Respiratory	X Diseases of the respiratory system
D	Digestive	XI Diseases of the digestive system
S	Skin	XII Diseases of the skin and subcutaneous tissue
L	Musculoskeletal	XIII Diseases of the musculoskeletal system and connective tissue
U&X&Y	Urology& Female genital system and breast & Male genital system	XIV Diseases of the genitourinary system
W	Pregnancy, childbirth, family planning	XV Pregnancy, childbirth and the puerperium
	NA	XVI Certain conditions originating in the perinatal period
	NA	XVII Congenital malformations, deformations and chromosomal abnormalities
A	General and unspecified	XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

	NA	XIX Injury, poisoning and certain other consequences of external causes
	NA	XX External causes of morbidity and mortality
Z	Social problems	XXI Factors influencing health status and contact with health services
	NA	XXII Codes for special purposes

Appendix B: Medical codes associated with antibiotic prescriptions in primary care in England

Table A 4: Medical codes associated with 99.8% antibiotic prescriptions in CPRD ranked descendent according to frequencies

Med code	Number of		Read term
	episodes	Read code	
16	116951	9N31.00	Telephone encounter
26	70142	6A...00	Patient reviewed
1	62477	246..00	O/E - blood pressure reading
92	50517	171..00	Cough
1273	47609	171..11	C/O - cough
18726	43161	9N3A.00	Telephone triage encounter
138	35355	H03..00	Acute tonsillitis
7579	34500	1J4..00	Suspected UTI
2581	33859	H06z000	Chest infection NOS
6154	28452	242..00	O/E - pulse rate
2	24969	22A..00	O/E - weight
68	23986	H06z011	Chest infection
5755	22571	1C9..00	Sore throat symptom
33	22156	1371.00	Never smoked tobacco
1289	20331	K190.00	Urinary tract infection, site not specified
7622	20013	8CAL.00	Smoking cessation advice
6039	19824	9N1C.11	Home visit
6677	19197	9Z...00	Administration NOS
38	18823	8CB..00	Had a chat to patient
980	18642	H01..00	Acute sinusitis
389	18054	K15..00	Cystitis
12200	17747	661M.00	Clinical management plan agreed
8034	17336	8B3S.00	Medication review
93	16522	137P.00	Cigarette smoker
90	16187	137S.00	Ex smoker
2637	15275	H05z.11	Upper respiratory tract infection NOS
74	14306	8B3H.00	Medication requested
76	13856	H05z.00	Upper respiratory infection NOS
292	13149	1719.00	Chesty cough
379	12865	M261000	Acne vulgaris
1649	12812	9....00	Administration

4207	12769	M03z000	Cellulitis NOS
4495	11919	M0...00	Skin and subcutaneous tissue infections
532	11424	1A55.00	Dysuria
57	11056	246..11	O/E - BP reading
6679	10338	2....11	Examination of patient
461	9772	8CA..00	Patient given advice
22811	9605	ZQ3J.00	Triage
667	9356	1A...12	Urinary symptoms
8797	8245	9N58.00	Emergency appointment
2963	8145	677B.00	Advice about treatment given
4703	7830	1D14.00	C/O: a rash
267	7810	F52z.00	Otitis media NOS
3358	7695	H06z100	Lower resp tract infection
8297	7272	Z4A..00	Discussion
1665	7111	8B31400	Medication review
			Urinary tract infection, site not specified
150	7066	K190z00	NOS
6573	6984	14L..00	H/O: drug allergy
11843	6895	8B3x.00	Medication review with patient
96	6890	81H..00	Dressing of wound
4556	6759	9N4..00	Failed encounter
17828	6268	23...00	Examn. of respiratory system
177	6193	1969.00	Abdominal pain
5813	6082	1C3..00	Earache symptoms
11955	6017	8B3V.00	Medication review done
312	5808	H060.00	Acute bronchitis
18645	5578	9N32.00	Third party encounter
2138	5563	F501.00	Infective otitis externa
3	5526	229..00	O/E - height
1683	5507	212..00	Patient examined
6373	5364	M07z.11	Infected insect bite
10043	5315	66YJ.00	Asthma annual review
9753	5191	1....00	History / symptoms
9513	4663	8CAZ.00	Patient given advice NOS
893	4444	H02..00	Acute pharyngitis
943	4376	M05..00	Impetigo
1135	4278	F587.11	Ear pain
9599	4276	1151.00	No known allergies
13174	4246	663Q.00	Asthma not limiting activities

			Advice about long acting reversible
96915	4246	8CAw.00	contraception
1367	4225	M244.00	Folliculitis
27	4205	136..00	Alcohol consumption
13173	4152	663O.00	Asthma not disturbing sleep
5754	4109	1BA5.11	Pain in sinuses
			Nonsuppurative otitis media + eustachian
5577	4085	F51..00	tube disorders
404	4024	1C9..11	Throat soreness
1474	3963	F52..00	Suppurative and unspecified otitis media
47	3931	679..11	Advice to patient - subject
			Acute exacerbation of chronic obstructive
1446	3903	H312200	airways disease
13380	3893	2128.00	Patient's condition the same
17690	3783	2....00	Examination / Signs
			Chronic obstructive pulmonary disease
11287	3690	66YM.00	annual review
185	3646	H333.00	Acute exacerbation of asthma
2651	3618	2D...00	Ear, nose + throat examination
2007	3582	M0z..11	Infected sebaceous cyst
6124	3557	H062.00	Acute lower respiratory tract infection
731	3533	F587.00	Otalgia
821	3498	M230.00	Ingrowing nail
25997	3494	8BP0.00	Deferred antibiotic therapy
5859	3459	1652.00	Feels hot/feverish
293	3430	H06z111	Respiratory tract infection
226	3339	1J...00	Suspected condition
35441	3279	9c0C.00	Result
19258	3243	9N3D.00	Letter received
5887	3172	F510.00	Acute non suppurative otitis media
17473	3114	9N7..12	Patient asked to come in
28910	3105	8BMC.00	Prescription collected by pharmacy
7917	3103	8H8..00	Follow-up arranged
1741	3059	M111.00	Atopic dermatitis/eczema
1160	3038	R062.00	[D]Cough
729	2980	1A1..13	Urinary frequency
3627	2922	1A7..00	Vaginal discharge symptom
4822	2898	1739.00	Shortness of breath
243	2827	H01..11	Sinusitis

			Failed encounter - message left on answer machine
6334	2811	9N4F.00	
1746	2803	16E..00	Feels unwell
6588	2788	ZV68100	[V]Issue of repeat prescription
17275	2716	9NE8.00	Fax sent to:
12949	2693	1361.00	Teetotaller
7877	2661	8HQ1.00	Refer for X-Ray
1103	2613	J025000	Dental abscess
5965	2568	M03z100	Abscess NOS
1356	2564	M01..00	Furuncle - boil
9378	2555	1AG..00	Recurrent urinary tract infections
3629	2534	8B31100	Medication given
48	2529	13A..00	Diet - patient initiated
18732	2521	8BC1.00	Treatment plan given
860	2471	2315.00	Resp. system examined - NAD
42824	2465	663q.00	Asthma daytime symptoms
140	2446	9N19.00	Seen in hospital casualty
1442	2436	M02z.12	Paronychia
36	2428	138..00	Exercise grading
			Brief intervention for physical activity completed
99138	2417	9Oq3.00	
6294	2361	H051.00	Acute upper respiratory tract infection
5926	2359	M230000	Ingrowing great toe nail
1139	2324	M2yz.11	Skin lesion
1763	2323	R090.00	[D]Abdominal pain
2364	2292	SP25500	Postoperative wound infection, unspecified
6366	2216	9N33.11	Letter encounter
4348	2186	F527.00	Acute right otitis media
7645	2184	9N4Z.00	Failed encounter NOS
5910	2141	M01z.00	Boil NOS
			Issue of chronic obstructive pulmonary disease rescue pack
101042	2132	8BMW.00	
10907	2129	9Na..00	Consultation
374	2083	182..00	Chest pain
26501	2065	663s.00	Asthma never causes daytime symptoms
14674	2054	8B41.00	Repeated prescription

Appendix C: Code lists for antibiotic prescription study

Table A 5: Read codes for respiratory conditions

Read code	Read term
1656	Feverish cold
1712	Dry cough
1713	Productive cough -clear sputum
1714	Productive cough -green sputum
1715	Productive cough-yellow sputum
1716	Productive cough NOS
1716.11	Coughing up phlegm
1717	Night cough present
1719	Chesty cough
1719.11	Bronchial cough
1739	Shortness of breath
14B2.00	H/O: pneumonia
14B3.11	H/O: bronchitis
14B9.00	History of acute lower respiratory tract infection
16L..00	Influenza-like symptoms
171..00	Cough
171..11	C/O - cough
171A.00	Chronic cough
171B.00	Persistent cough
171C.00	Morning cough
171D.00	Evening cough
171E.00	Unexplained cough
171F.00	Cough with fever
171G.00	Bovine cough
171H.00	Difficulty in coughing up sputum
171J.00	Reflux cough
171K.00	Barking cough
171Z.00	Cough symptom NOS
173..00	Breathlessness
173B.00	Nocturnal cough / wheeze
1BA5.11	Pain in sinuses
1c3..00	Earache symptoms
1C3..00	Earache symptoms

1C32.00	Unilateral earache
1C33.00	Bilateral earache
1C3Z.00	Earache symptom NOS
1C9..00	Sore throat symptom
1C9..11	Throat soreness
1C92.00	Has a sore throat
1C93.00	Persistent sore throat
1C9Z.00	Sore throat symptom NOS
1CB..00	Throat symptom NOS
1CB3.00	Throat pain
1CB3.11	Pain in throat
1CB4.00	Feeling of lump in throat
1CB4.11	Constriction in throat
1CB4.12	Tightness in throat
1CB5.00	Throat irritation
1CBZ.00	Throat symptom NOS
2DB6.00	O/E - follicular tonsillitis
2DC2.00	O/E - granular pharyngitis
2DC3.00	Inflamed throat
A022200	Salmonella pneumonia
A32..00	Diphtheria
A320.00	Faucial diphtheria
A321.00	Nasopharyngeal diphtheria
A322.00	Anterior nasal diphtheria
A323.00	Laryngeal diphtheria
A32y.00	Other specified diphtheria
A32y000	Conjunctival diphtheria
A32y400	Cutaneous diphtheria
A32yz00	Other specified diphtheria NOS
A32z.00	Diphtheria NOS
A33..00	Whooping cough
A33..11	Bordetella
A330.00	Bordetella pertussis
A331.00	Bordetella parapertussis
A33y.00	Whooping cough - other specified organism
A33y000	Bordetella bronchiseptica
A33yz00	Other whooping cough NOS
A33z.00	Whooping cough NOS

A34..00	Streptococcal sore throat and scarlatina
A340.00	Streptococcal sore throat
A340100	Streptococcal laryngitis
A340200	Streptococcal pharyngitis
A340300	Streptococcal tonsillitis
A340z00	Streptococcal sore throat NOS
A34z.00	Streptococcal sore throat with scarlatina NOS
A383000	Fusobacterial necrotising tonsillitis
A54x400	Herpes simplex pneumonia
A551.00	Postmeasles pneumonia
A552.00	Postmeasles otitis media
A730.00	Ornithosis with pneumonia
A789300	HIV disease resulting in Pneumocystis carinii pneumonia
A789311	HIV disease resulting in Pneumocystis jirovecii pneumonia
AA12.00	Vincent's pharyngitis
AA1z.11	Vincent's laryngitis
AA1z.12	Vincent's tonsillitis
AA25.11	Rhinopharyngitis mutilans
AB24.11	Pneumonia - candidal
AB40500	Histoplasma capsulatum with pneumonia
AB41500	Histoplasma duboisii with pneumonia
AyuK900	[X]Mycoplasma pneumoniae [PPLO]cause/dis classifd/other chapter
AyuKA00	[X]Klebsiella pneumoniae/cause/disease classifd/other chapters
F00y400	Meningitis due to klebsiella pneumoniae
F501.00	Infective otitis externa
F51..00	Nonsuppurative otitis media + eustachian tube disorders
F510.00	Acute non suppurative otitis media
F510000	Acute otitis media with effusion
F510011	Acute secretory otitis media
F510100	Acute serous otitis media
F510200	Acute mucoid otitis media
F510300	Acute sanguinous otitis media
F510z00	Acute nonsuppurative otitis media NOS
F514.00	Unspecified nonsuppurative otitis media
F514100	Serous otitis media NOS
F514200	Catarrhal otitis media NOS

F514300	Mucoid otitis media NOS
F514z00	Nonsuppurative otitis media NOS
F52..00	Suppurative and unspecified otitis media
F520.00	Acute suppurative otitis media
F520000	Acute suppurative otitis media tympanic membrane intact
F520100	Acute suppurative otitis media tympanic membrane ruptured
F520300	Acute suppurative otitis media due to disease EC
F520z00	Acute suppurative otitis media NOS
F521.00	Chronic suppurative otitis media, tubotympanic
F522.00	Chronic suppurative otitis media, atticofurcal
F523.00	Chronic suppurative otitis media NOS
F524.00	Purulent otitis media NOS
F524000	Bilateral suppurative otitis media
F525.00	Recurrent acute otitis media
F526.00	Acute left otitis media
F527.00	Acute right otitis media
F528.00	Acute bilateral otitis media
F52z.00	Otitis media NOS
F52z.11	Infection ear
F53..00	Mastoiditis and related conditions
F530.00	Acute mastoiditis
F530.11	Abscess of mastoid
F530.12	Empyema of mastoid
F530000	Acute mastoiditis without complications
F530100	Subperiosteal mastoid abscess
F530200	Gradenigo's syndrome
F530300	Acute mastoiditis with other complication
F530z00	Acute mastoiditis NOS
F531.00	Chronic mastoiditis
F531000	Caries of mastoid
F531100	Post aural mastoid fistula
F531z00	Chronic mastoiditis NOS
F532.00	Petrositis
F5329	Petrositis
F533.00	Postmastoidectomy complication
F533000	Unspecified postmastoidectomy complication
F533100	Postmastoidectomy cavity mucinous cyst
F533200	Recurrent cholesteatoma postmastoidectomy

F533300	Postmastoidectomy granulation cavity
F533z00	Postmastoidectomy complication NOS
F53y.00	Other mastoid disorders
F53y000	Postauricular fistula
F53y100	Other mastoid disorder NOS
F53z.00	Mastoiditis NOS
F540.00	Acute myringitis without otitis media
F540z00	Acute myringitis NOS
F587.00	Otalgia
F587.11	Ear pain
FyuP000	[X]Other acute nonsuppurative otitis media
FyuP200	[X]Other chronic suppurative otitis media
FyuP300	[X]Otitis media in bacterial diseases classified elsewhere
FyuP400	[X]Otitis media in viral diseases classified elsewhere
FyuP500	[X]Otitis media in other diseases classified elsewhere
H0...00	Acute respiratory infections
H00..00	Acute nasopharyngitis
H00..11	Common cold
H00..12	Coryza - acute
H00..13	Febrile cold
H00..14	Nasal catarrh - acute
H00..15	Pyrexial cold
H00..16	Rhinitis - acute
H01..00	Acute sinusitis
H01..11	Sinusitis
H010.00	Acute maxillary sinusitis
H010.11	Antritis - acute
H011.00	Acute frontal sinusitis
H012.00	Acute ethmoidal sinusitis
H013.00	Acute sphenoidal sinusitis
H014.00	Acute rhinosinusitis
H01y.00	Other acute sinusitis
H01y000	Acute pansinusitis
H01yz00	Other acute sinusitis NOS
H01z.00	Acute sinusitis NOS
H02..00	Acute pharyngitis
H02..11	Sore throat NOS
H02..12	Viral sore throat NOS

H02..13	Throat infection - pharyngitis
H020.00	Acute gangrenous pharyngitis
H021.00	Acute phlegmonous pharyngitis
H022.00	Acute ulcerative pharyngitis
H023.00	Acute bacterial pharyngitis
H023000	Acute pneumococcal pharyngitis
H023100	Acute staphylococcal pharyngitis
H023z00	Acute bacterial pharyngitis NOS
H024.00	Acute viral pharyngitis
H02z.00	Acute pharyngitis NOS
H03..00	Acute tonsillitis
H03..11	Throat infection - tonsillitis
H03..12	Tonsillitis
H030.00	Acute erythematous tonsillitis
H031.00	Acute follicular tonsillitis
H032.00	Acute ulcerative tonsillitis
H033.00	Acute catarrhal tonsillitis
H034.00	Acute gangrenous tonsillitis
H035.00	Acute bacterial tonsillitis
H035000	Acute pneumococcal tonsillitis
H035100	Acute staphylococcal tonsillitis
H035z00	Acute bacterial tonsillitis NOS
H036.00	Acute viral tonsillitis
H037.00	Recurrent acute tonsillitis
H03z.00	Acute tonsillitis NOS
H04..00	Acute laryngitis and tracheitis
H040.00	Acute laryngitis
H040000	Acute oedematous laryngitis
H040100	Acute ulcerative laryngitis
H040200	Acute catarrhal laryngitis
H040300	Acute phlegmonous laryngitis
H040400	Acute haemophilus influenzae laryngitis
H040600	Acute suppurative laryngitis
H040w00	Acute viral laryngitis unspecified
H040x00	Acute bacterial laryngitis unspecified
H040z00	Acute laryngitis NOS
H041.00	Acute tracheitis
H041000	Acute tracheitis without obstruction

H041100	Acute tracheitis with obstruction
H041z00	Acute tracheitis NOS
H042.00	Acute laryngotracheitis
H042.11	Laryngotracheitis
H042000	Acute laryngotracheitis without obstruction
H042100	Acute laryngotracheitis with obstruction
H042z00	Acute laryngotracheitis NOS
H043.00	Acute epiglottitis (non strep)
H043.11	Viral epiglottitis
H043000	Acute epiglottitis without obstruction
H043100	Acute epiglottitis with obstruction
H043200	Acute obstructive laryngitis
H043211	Croup
H043z00	Acute epiglottitis NOS
H044.00	Croup
H04z.00	Acute laryngitis and tracheitis NOS
H05..00	Other acute upper respiratory infections
H050.00	Acute laryngopharyngitis
H051.00	Acute upper respiratory tract infection
H052.00	Pharyngotracheitis
H053.00	Tracheopharyngitis
H054.00	Recurrent upper respiratory tract infection
H055.00	Pharyngolaryngitis
H05y.00	Other upper respiratory infections of multiple sites
H05z.00	Upper respiratory infection NOS
H05z.11	Upper respiratory tract infection NOS
H05z.12	Viral upper respiratory tract infection NOS
H06..00	Acute bronchitis and bronchiolitis
H060.00	Acute bronchitis
H060.11	Acute wheezy bronchitis
H060000	Acute fibrinous bronchitis
H060100	Acute membranous bronchitis
H060200	Acute pseudomembranous bronchitis
H060300	Acute purulent bronchitis
H060400	Acute croupous bronchitis
H060500	Acute tracheobronchitis
H060600	Acute pneumococcal bronchitis
H060700	Acute streptococcal bronchitis

H060800	Acute haemophilus influenzae bronchitis
H060900	Acute neisseria catarrhalis bronchitis
H060A00	Acute bronchitis due to mycoplasma pneumoniae
H060B00	Acute bronchitis due to coxsackievirus
H060C00	Acute bronchitis due to parainfluenza virus
H060D00	Acute bronchitis due to respiratory syncytial virus
H060E00	Acute bronchitis due to rhinovirus
H060F00	Acute bronchitis due to echovirus
H060v00	Subacute bronchitis unspecified
H060w00	Acute viral bronchitis unspecified
H060x00	Acute bacterial bronchitis unspecified
H060z00	Acute bronchitis NOS
H061.00	Acute bronchiolitis
H061000	Acute capillary bronchiolitis
H061100	Acute obliterating bronchiolitis
H061200	Acute bronchiolitis with bronchospasm
H061300	Acute exudative bronchiolitis
H061500	Acute bronchiolitis due to respiratory syncytial virus
H061600	Acute bronchiolitis due to other specified organisms
H061z00	Acute bronchiolitis NOS
H062.00	Acute lower respiratory tract infection
H06z.00	Acute bronchitis or bronchiolitis NOS
H06z000	Chest infection NOS
H06z011	Chest infection
H06z100	Lower resp tract infection
H06z111	Respiratory tract infection
H06z112	Acute lower respiratory tract infection
H06z200	Recurrent chest infection
H07..00	Chest cold
H0y..00	Other specified acute respiratory infections
H0z..00	Acute respiratory infection NOS
H121100	Atrophic pharyngitis
H121200	Granular pharyngitis
H121300	Hypertrophic pharyngitis
H121400	Pharyngitis keratosa
H130.12	Maxillary sinusitis
H131.11	Frontal sinusitis
H135.00	Recurrent sinusitis

H13y100	Pansinusitis
H14y600	Lingular tonsillitis
H2...00	Pneumonia and influenza
H20..00	Viral pneumonia
H20..11	Chest infection - viral pneumonia
H200.00	Pneumonia due to adenovirus
H201.00	Pneumonia due to respiratory syncytial virus
H202.00	Pneumonia due to parainfluenza virus
H203.00	Pneumonia due to human metapneumovirus
H20y.00	Viral pneumonia NEC
H20y000	Severe acute respiratory syndrome
H20z.00	Viral pneumonia NOS
H21..00	Lobar (pneumococcal) pneumonia
H21..11	Chest infection - pneumococcal pneumonia
H22..00	Other bacterial pneumonia
H22..11	Chest infection - other bacterial pneumonia
H220.00	Pneumonia due to klebsiella pneumoniae
H221.00	Pneumonia due to pseudomonas
H222.00	Pneumonia due to haemophilus influenzae
H222.11	Pneumonia due to haemophilus influenzae
H223.00	Pneumonia due to streptococcus
H223000	Pneumonia due to streptococcus, group B
H224.00	Pneumonia due to staphylococcus
H22y.00	Pneumonia due to other specified bacteria
H22y000	Pneumonia due to escherichia coli
H22y011	E.coli pneumonia
H22y100	Pneumonia due to proteus
H22y200	Pneumonia - Legionella
H22yX00	Pneumonia due to other aerobic gram-negative bacteria
H22yz00	Pneumonia due to bacteria NOS
H22z.00	Bacterial pneumonia NOS
H23..00	Pneumonia due to other specified organisms
H23..11	Chest infection - pneumonia organism OS
H230.00	Pneumonia due to Eaton's agent
H231.00	Pneumonia due to mycoplasma pneumoniae
H232.00	Pneumonia due to pleuropneumonia like organisms
H233.00	Chlamydial pneumonia
H23z.00	Pneumonia due to specified organism NOS

H24..00	Pneumonia with infectious diseases EC
H24..11	Chest infection with infectious disease EC
H240.00	Pneumonia with measles
H241.00	Pneumonia with cytomegalic inclusion disease
H242.00	Pneumonia with ornithosis
H243.00	Pneumonia with whooping cough
H243.11	Pneumonia with pertussis
H244.00	Pneumonia with tularaemia
H246.00	Pneumonia with aspergillosis
H247000	Pneumonia with candidiasis
H247100	Pneumonia with coccidioidomycosis
H247z00	Pneumonia with systemic mycosis NOS
H24y.00	Pneumonia with other infectious diseases EC
H24y000	Pneumonia with actinomycosis
H24y100	Pneumonia with nocardiasis
H24y200	Pneumonia with pneumocystis carinii
H24y300	Pneumonia with Q-fever
H24y400	Pneumonia with salmonellosis
H24y500	Pneumonia with toxoplasmosis
H24y600	Pneumonia with typhoid fever
H24y700	Pneumonia with varicella
H24yz00	Pneumonia with other infectious diseases EC NOS
H24z.00	Pneumonia with infectious diseases EC NOS
H25..00	Bronchopneumonia due to unspecified organism
H25..11	Chest infection - unspecified bronchopneumonia
H26..00	Pneumonia due to unspecified organism
H26..11	Chest infection - pneumonia due to unspecified organism
H260.00	Lobar pneumonia due to unspecified organism
H260000	Lung consolidation
H261.00	Basal pneumonia due to unspecified organism
H262.00	Postoperative pneumonia
H263.00	Pneumonitis, unspecified
H27..00	Influenza
H270.00	Influenza with pneumonia
H270.11	Chest infection - influenza with pneumonia
H270000	Influenza with bronchopneumonia
H270100	Influenza with pneumonia, influenza virus identified
H270z00	Influenza with pneumonia NOS

H271.00	Influenza with other respiratory manifestation
H271000	Influenza with laryngitis
H271100	Influenza with pharyngitis
H271z00	Influenza with respiratory manifestations NOS
H27y.00	Influenza with other manifestations
H27y100	Influenza with gastrointestinal tract involvement
H27yz00	Influenza with other manifestations NOS
H27z.00	Influenza NOS
H27z.11	Flu like illness
H27z.12	Influenza like illness
H28..00	Atypical pneumonia
H29..00	Avian influenza
H2A..00	Influenza due to Influenza A virus subtype H1N1
H2A..11	Influenza A (H1N1) swine flu
H2B..00	Community acquired pneumonia
H2C..00	Hospital acquired pneumonia
H2y..00	Other specified pneumonia or influenza
H2z..00	Pneumonia or influenza NOS
H30..00	Bronchitis unspecified
H30..11	Chest infection - unspecified bronchitis
H300.00	Tracheobronchitis NOS
H301.00	Laryngotracheobronchitis
H30z.00	Bronchitis NOS
H310100	Smokers' cough
H312200	Acute exacerbation of chronic obstructive airways disease
H470312	Aspiration pneumonia
H471000	Aspiration pneumonia due to vomit
H50..00	Lipoid pneumonia (exogenous)
H500.00	Empyema
H500000	Empyema with fistula
H500100	Empyema with bronchocutaneous fistula
H500400	Empyema with bronchopleural fistula
H501.00	Empyema with pleural fistula NOS
H501000	Empyema with no fistula
H501100	Pleural abscess
H501200	Thorax abscess NOS
H501300	Pleural empyema
H501400	Lung empyema NOS

H501500	Purulent pleurisy
H501600	Pyopneumothorax
H50z.00	Pyothorax
H51..00	Empyema NOS
H510.00	Pleurisy
H510000	Pleurisy without effusion or active tuberculosis
H510100	Adhesion of pleura or lung
H510200	Thickening of pleura
H510300	Calcification of pleura
H510400	Acute dry pleurisy
H510500	Diaphragmatic pleurisy
H510600	Basal pleurisy
H510700	Chronic dry pleurisy
H510800	Fibrinous pleurisy
H510900	Sterile pleurisy
H510A00	Pneumococcal pleurisy
H510B00	Staphylococcal pleurisy
H510C00	Streptococcal pleurisy
H510z00	Pleural plaque
H511.00	Pleurisy without effusion or active tuberculosis NOS
H511000	Bacterial pleurisy with effusion
H511100	Pneumococcal pleurisy with effusion
H511200	Staphylococcal pleurisy with effusion
H511z00	Streptococcal pleurisy with effusion
H51y.00	Bacterial pleurisy with effusion NOS
H51y000	Other pleural effusion excluding mention of tuberculosis
H51y100	Encysted pleurisy
H51y200	Haemopneumothorax
H51y300	Haemothorax
H51y400	Hydropneumothorax
H51y500	Hydrothorax
H51y600	Chylous effusion
H51y700	Fibrothorax
H51yz00	Malignant pleural effusion
H51z.00	Other pleural effusion
H51z000	Pleural effusion NOS
H51z100	Exudative pleurisy NOS
H51z200	Serofibrinous pleurisy NOS

H51zz00	Serous pleurisy NOS
H530200	Pleural effusion NOS
H530300	Gangrenous pneumonia
H540000	Abscess of lung with pneumonia
H540100	Hypostatic pneumonia
Hyu0.00	Hypostatic bronchopneumonia
Hyu0000	[X]Acute upper respiratory infections
Hyu0100	[X]Other acute sinusitis
Hyu0200	[X]Acute pharyngitis due to other specified organisms
Hyu0300	[X]Other acute upper respiratory infections/multiple sites
Hyu0500	[X]Influenza+other manifestations,influenza virus identified
Hyu0600	[X]Influenza+oth respiratory manifestatns,virus not identified
Hyu0700	[X]Influenza+other manifestations, virus not identified
Hyu0800	[X]Other viral pneumonia
Hyu0A00	[X]Other bacterial pneumonia
Hyu0B00	[X]Pneumonia due to other specified infectious organisms
Hyu0D00	[X]Pneumonia in viral diseases classified elsewhere
Hyu0H00	[X]Other pneumonia, organism unspecified
Hyu1.00	[X]Other acute lower respiratory infections
Hyu1000	[X]Acute bronchitis due to other specified organisms
Hyu1100	[X]Acute bronchiolitis due to other specified organisms
M03z000	Cellulitis NOS
R041.00	[D]Throat pain
R062.00	[D]Cough
SN30.11	Aero-otitis media
SN31.11	Aerosinusitis
SP13100	Other aspiration pneumonia as a complication of care
SP13200	Post operative chest infection

Table A 6: Read codes for genitourinary conditions.

Read code	Read term
1979	Suprapubic pain
14D2.00	H/O: kidney infection
14D4.00	H/O: recurrent cystitis
14D5.00	H/O: haematuria
14D6.00	H/O: urethral stricture
14D7.00	History of recurrent urinary tract infection
14DZ.00	H/O: urinary disease NOS
16F..00	Double incontinence
1A...00	Genitourinary symptoms
1A...11	GU symptoms
1A...12	Urinary symptoms
1A1..00	Micturition frequency
1A1..11	Frequency of micturition
1A1..12	Polyuria
1A1..13	Urinary frequency
1A12.00	Frequency of micturition
1A13.00	Nocturia
1A1Z.00	Micturition frequency NOS
1A2..00	Micturition control
1A2..11	Urinary control
1A22.00	Enuresis
1A22000	Nocturnal enuresis
1A22011	Bedwetting
1A22100	Daytime enuresis
1A23.00	Incontinence of urine
1A24.00	Stress incontinence
1A24.11	Stress incontinence - symptom
1A25.00	Urgency
1A25.11	Urgency of micturition
1A26.00	Urge incontinence of urine
1A27.00	Urge to pass urine again shortly after finishing voiding
1A2Z.00	Micturition control NOS
1A3..00	Micturition stream
1A3..11	Urine stream
1A32.00	Cannot pass urine - retention
1A32.11	Retention - symptom

1A33.00	Micturition stream poor
1A34.00	Hesitancy
1A34.11	Hesitancy of micturition
1A35.00	Precipitancy
1A35.11	Precipitancy of micturition
1A36.00	Terminal dribbling of urine
1A37.00	Dribbling of urine
1A3Z.00	Micturition stream NOS
1A4..00	Urine appearance
1A4..11	Urine appearance symptom
1A41.00	Urine looks normal
1A42.00	Urine looks dark
1A43.00	Urine looks pale
1A44.00	Urine looks cloudy
1A45.00	Blood in urine - haematuria
1A45.11	Blood in urine - symptom
1A45.12	Haematuria - symptom
1A4Z.00	Urine appearance NOS
1A5..00	Genitourinary pain
1A51.00	No genitourinary pain
1A52.00	Renal colic
1A52.11	Renal colic, symptom
1A53.00	Lumbar ache - renal
1A53.11	C/O - loin pain
1A53.12	C/O - lumbar pain
1A53.13	C/O - renal pain
1A54.00	Ureteric colic
1A54.11	C/O - ureteric colic
1A54.12	C/O - ureteric pain
1A55.00	Dysuria
1A56.00	Strangury
1A57.00	Pain in testicle
1A57.11	Testicular pain
1A5B.00	Pain in penis
1A5D.00	Urethral pain
1A5Z.00	Genitourinary pain NOS
1A6..00	Urethral discharge symptom
1A61.00	No urethral discharge
1A62.00	Urethral discharge

1A6Z.00	Urethral discharge NOS
1AC..00	Micturition volume
1AC0.00	Anuria
1AC1.00	Oligouria
1AC2.00	Polyuria
1AF..00	Diuresis
1AG..00	Recurrent urinary tract infections
1AH..00	Bladder emptying
1AH0.00	Incomplete emptying of bladder
1AZ..00	Genitourinary symptoms NOS
1AZ3.00	Difficulty with micturition
1AZ6.00	Lower urinary tract symptoms
1AZ6000	Mild lower urinary tract symptoms
1AZ6100	Moderate lower urinary tract symptoms
1AZ6200	Severe lower urinary tract symptoms
1AZZ.00	Genitourinary symptom NOS
1J4..00	Suspected UTI
46B..00	Urine bacteriuria test
46B3.00	Urine bacteria test: positive
46B4.00	Urinary pneumococcal antigen test
46BZ.00	Urine bacteria test NOS
46f..00	Urine leucocyte test
46f2.00	Urine leucocyte test = +
46f3.00	Urine leucocyte test = ++
46f4.00	Urine leucocyte test = +++
46f5.00	Urine leucocyte test = trace
46G..00	Urine microscopy: cells
46G4.11	Leucocytes in urine
46G4.12	Sterile pyuria
46G5.00	Urine micr.: leucs - % polys
46G8.00	Urine Microscopy: white cells
46GZ.00	Urine microscopy: cells NOS
K1...00	Other urinary system diseases
K10..00	Infections of kidney
K10..11	Renal infections
K100.00	Chronic pyelonephritis
K100000	Chronic pyelonephritis without medullary necrosis
K100100	Chronic pyelonephritis with medullary necrosis
K100200	Chronic pyelitis

K100300	Chronic pyonephrosis
K100400	Nonobstructive reflux-associated chronic pyelonephritis
K100500	Chronic obstructive pyelonephritis
K100600	Calculous pyelonephritis
K100z00	Chronic pyelonephritis NOS
K101.00	Acute pyelonephritis
K101000	Acute pyelonephritis without medullary necrosis
K101200	Acute pyelitis
K101300	Acute pyonephrosis
K101z00	Acute pyelonephritis NOS
K102.00	Renal and perinephric abscess
K102000	Renal abscess
K102100	Perinephric abscess
K102200	Renal carbuncle
K102z00	Renal and perinephric abscess NOS
K103.00	Pyeloureteritis cystica
K103.11	Ureteritis cystica
K103.12	Infestation of renal pelvis with ureter
K104.00	Xanthogranulomatous pyelonephritis
K105.00	Chronic infective interstitial nephritis
K106.00	Candida pyelonephritis
K10y.00	Pyelonephritis and pyonephrosis unspecified
K10y000	Pyelonephritis unspecified
K10y100	Pyelitis unspecified
K10y200	Pyonephrosis unspecified
K10y300	Pyelonephritis in diseases EC
K10y400	Pyelitis in diseases EC
K10yz00	Unspecified pyelonephritis NOS
K10z.00	Infection of kidney NOS
K112.00	Hydronephrosis with renal and ureteral calculous obstruction
K113.00	Hydronephrosis with ureteropelvic junction obstruction
K113.11	Hydronephrosis with pelviureteric junction obstruction
K11X.00	Hydronephrosis with ureteral stricture NEC
K11z.00	Hydronephrosis NOS
K12..00	Calculus of kidney and ureter
K12..11	Kidney calculus
K12..12	Urinary calculus
K120.00	Calculus of kidney
K120.11	Nephrolithiasis NOS

K120.12	Renal calculus
K120.13	Renal stone
K120000	Staghorn calculus
K120z00	Renal calculus NOS
K121.00	Calculus of ureter
K121.11	Ureteric calculus
K121.12	Ureteric stone
K121.13	Ureterolithiasis
K122.00	Calculus of kidney with calculus of ureter
K12z.00	Urinary calculus NOS
K13..00	Other kidney and ureter disorders
K13..11	Other kidney disorders
K13..12	Other ureter disorders
K132.00	Acquired cyst of kidney
K132.11	Acquired renal cystic disease
K132000	Single acquired kidney cyst
K132100	Multiple acquired kidney cysts
K132200	Peripelvic (lymphatic) cyst
K132300	Acquired renal cyst with neoplastic change
K132400	Acquired renal cyst without neoplastic change
K132z00	Acquired cyst of kidney NOS
K133.00	Stricture of ureter
K133000	Postoperative ureteric constriction
K133100	Stricture of pelviureteric junction
K133z00	Stricture of ureter NOS
K134.00	Other ureteric obstruction
K134z00	Occlusion of ureter NOS
K135.00	Hydroureter
K136.11	Orthostatic proteinuria
K137.00	Vesicoureteric reflux
K137.11	Ureteric reflux
K13B.00	Calyceal diverticulum
K13y600	Ureterocele - acquired
K13y611	Idiopathic dilation of ureter
K13y700	Megaloureter - acquired
K13y800	Perirenal haematoma
K13y900	Ureteric neuromuscular incoordination
K13yA00	Dent's disease
K13yz00	Other kidney and ureteric disorders NOS

K13z.00	Kidney and ureter disease NOS
K14..00	Lower urinary tract calculus
K140.00	Bladder calculus
K140.11	Bladder stone
K140000	Calculus in diverticulum of bladder
K140100	Other calculus in bladder
K140z00	Bladder calculus NOS
K141.00	Calculus in urethra
K14y.00	Other lower urinary tract calculus
K14z.00	Lower urinary tract calculus NOS
K15..00	Cystitis
K150.00	Acute cystitis
K151.00	Chronic interstitial cystitis
K151000	Hunner's ulcer
K151100	Panmural fibrosis of bladder
K151200	Submucous cystitis
K151z00	Chronic interstitial cystitis NOS
K152.00	Other chronic cystitis
K152000	Subacute cystitis
K152y00	Chronic cystitis unspecified
K152z00	Other chronic cystitis NOS
K153.00	Trigonitis
K153.11	Follicular cystitis
K153000	Acute trigonitis
K153100	Chronic trigonitis
K153200	Urethrotigonitis
K153z00	Trigonitis NOS
K154.00	Cystitis in diseases EC
K154000	Cystitis in actinomycosis
K154100	Cystitis in amoebiasis
K154200	Cystitis in bilharziasis
K154500	Cystitis in gonorrhoea
K154600	Cystitis in moniliasis
K154700	Cystitis in trichomoniasis
K154800	Cystitis in tuberculosis
K154z00	Cystitis in diseases EC NOS
K155.00	Recurrent cystitis
K15y.00	Other specified cystitis
K15y000	Cystitis cystica

K15y100	Irradiation cystitis
K15y200	Abscess of bladder
K15y300	Malakoplakia of bladder
K15yz00	Other cystitis NOS
K15z.00	Cystitis NOS
K16..00	Other disorders of bladder
K160.00	Bladder neck obstruction
K160.11	Contracture of bladder neck
K160.12	Stenosis of bladder neck
K160.13	BOO - Bladder outflow obstruction
K161.00	Intestino-vesical fistula
K161000	Enterovesical fistula
K161100	Vesicocolic fistula
K161111	Colovesical fistula
K161200	Vesicosigmoidal fistula
K161300	Vesicorectal fistula
K161z00	Intestino-vesical fistula NOS
K162.00	Vesical fistula NEC
K162000	Vesicocutaneous fistula
K162100	Vesicoperineal fistula
K162200	Urethrovesical fistula
K162z00	Vesical fistula NEC NOS
K163.00	Diverticulum of bladder
K163000	Acquired bladder diverticulum
K163100	False bladder diverticulum
K163200	Bladder diverticulitis
K163z00	Diverticulum of bladder NOS
K164.00	Atony of bladder
K164.11	Atonic bladder
K164000	Hypotonic bladder
K164100	Bladder inertia
K164z00	Atony of bladder NOS
K165.00	Other functional disorder of bladder
K165000	Hypertonic bladder sphincter
K165100	Bladder sphincter paralysis
K165200	Bladder outflow obstruction
K165300	Detrusor instability
K165400	Unstable bladder
K165z00	Other bladder function disorder NOS

K166.00	Bladder rupture due to nontraumatic cause
K167.00	Haemorrhage into bladder wall
K168.00	Amyloid of bladder
K16V.00	Neuromuscular dysfunction of bladder, unspecified
K16V000	Neuropathic bladder
K16V011	Neurogenic bladder
K16V100	Overactive bladder
K16W.00	Reflex neuropathic bladder, not elsewhere classified
K16X.00	Uninhibited neuropathic bladder, NEC
K16y.00	Other bladder disorders
K16y000	Calcified bladder
K16y100	Contracted bladder
K16y200	Bladder haemorrhage
K16y300	Bladder hypertrophy
K16y400	Irritable bladder
K16y411	Detrusor instability
K16y412	Unstable bladder
K16y500	Trabeculation of bladder
K16y700	Squamous metaplasia of bladder
K16y800	Functional disorder of bladder
K16y811	Functional voiding disorder
K16y900	Metaplasia of trigone
K16yA00	Bladder scarring
K16yz00	Other bladder disorders NOS
K16z.00	Bladder disorders NOS
K17..00	Urethritis due to non venereal causes
K17..11	Periurethritis
K170.00	Urethral and periurethral abscess
K170.11	Urethral abscess
K170000	Urethral abscess unspecified
K170100	Bulbourethral gland abscess
K170111	Cowper's gland abscess
K170200	Urethral gland abscess
K170300	Periurethral cellulitis
K170311	Periurethritis
K170400	Periurethral abscess
K170z00	Urethral abscess NOS
K171.00	Post menopausal atrophic urethritis
K171.11	Post menopausal urethritis

K172.00	Candidal urethritis
K17y.00	Other urethritis
K17y000	Urethritis unspecified
K17y100	Urethral syndrome NOS
K17y200	Skene's glands adenitis
K17y300	Cowperitis
K17y400	Urethral meatitis
K17y500	Urethral meatal ulcer
K17y600	Verumontanitis
K17y700	Utricle masculinus
K17yz00	Other urethritis NOS
K17z.00	Urethritis due to non venereal cause NOS
K18..00	Urethral stricture
K18..11	Pinhole meatus
K180.00	Infective urethral stricture
K180000	Urethral stricture due to unspecified infection
K180100	Urethral stricture due to infection EC
K180z00	Infective urethral stricture NOS
K181.00	Traumatic urethral stricture
K181.11	Postobstetric urethral stricture
K182.00	Postoperative urethral stricture
K182.11	Postcatheterisation urethral stricture
K183.00	Stenosis of urinary meatus
K18y.00	Other urethral stricture
K18z.00	Urethral stricture NOS
K19..00	Other urethral and urinary tract disorders
K19..11	Other urethral disorders
K190.00	Urinary tract infection, site not specified
K190.11	Recurrent urinary tract infection
K190000	Bacteriuria, site not specified
K190011	Asymptomatic bacteriuria
K190100	Pyuria, site not specified
K190200	Post operative urinary tract infection
K190300	Recurrent urinary tract infection
K190311	Recurrent UTI
K190400	Chronic urinary tract infection
K190500	Urinary tract infection
K190600	Urosepsis
K190X00	Persistent proteinuria, unspecified

K190z00	Urinary tract infection, site not specified NOS
K191.00	Urethral fistula
K191000	Urethroperineal fistula
K191100	Urethrorectal fistula
K191z00	Urethral fistula NOS
K192.00	Urethral diverticulum
K193.00	Urethral caruncle
K193.11	Urethral polyp
K194.00	Urethral false passage
K195.00	Prolapsed urethral mucosa
K195.11	Urethrocele
K196.00	Urinary obstruction unspecified
K196.11	Obstructive uropathy, unspecified
K197.00	Haematuria
K197.11	Traumatic haematuria
K197.12	Essential haematuria
K197000	Painless haematuria
K197100	Painful haematuria
K197200	Microscopic haematuria
K197300	Frank haematuria
K197400	Clot haematuria
K197500	Benign familial haematuria
K198.00	Stress incontinence
K19C.00	Other obstructive and reflux uropathy
K19W.00	Urethral disorder, unspecified
K19X.00	Obstructive and reflux uropathy, unspecified
K19y.00	Other urinary tract disorders
K19y000	Urethral rupture due to nontraumatic cause
K19y100	Urethral cyst
K19y200	Urethral granuloma
K19y300	Pneumaturia
K19y400	Bleeding from urethra
K19y411	Urethral bleeding
K19yw00	Disorder of urinary system, unspecified
K19yz00	Other urinary tract disorders NOS
K19z.00	Urethral and urinary tract disorders NOS
K1A..00	Urinary calculus in schistosomiasis
K1y..00	Other specified diseases of urinary system
K1z..00	Other urinary system diseases NOS

Ky...00	Other specified diseases of genitourinary system
Kyu..00	[X]Additional genitourinary disease classification terms
Kyu1200	[X]Other obstructive and reflux uropathy
Kyu1300	[X]Obstructive and reflux uropathy, unspecified
Kyu1F00	[X]Hydronephrosis with ureteral stricture NEC
Kyu3.00	[X]Urolithiasis
Kyu3000	[X]Other lower urinary tract calculus
Kyu3100	[X]Calculus of urinary tract in other diseases CE
Kyu4.00	[X]Other disorders of kidney and ureter
Kyu4100	[X]Other specified disorders of kidney and ureter
Kyu5.00	[X]Other diseases of urinary system
Kyu5000	[X]Other chronic cystitis
Kyu5100	[X]Other cystitis
Kyu5200	[X]Other neuromuscular dysfunction of bladder
Kyu5300	[X]Other specified disorders of bladder
Kyu5500	[X]Other urethritis
Kyu5600	[X]Other urethral stricture
Kyu5A00	[X]Other specified urinary incontinence
Kyu5B00	[X]Other specified disorders of urinary system
Kyu5E00	[X]Neuromuscular dysfunction of bladder, unspecified
Kyu5F00	[X]Urethral disorder, unspecified
Kyu6100	[X]Other specified disorders of prostate
Kyu8200	[X]Other diseases of Bartholin's gland
Kyu8800	[X]Disease of Bartholin's gland, unspecified
Kyu9200	[X]Other female urinary-genital tract fistulae
Kyu9300	[X]Other female intestinal-genital tract fistulae
Kyu9400	[X]Other female genital tract fistulae
Kyu9F00	perimenopausal disorders
KyuA.00	[X]Other disorders of genitourinary tract
KyuA000	[X]Other postprocedural disorders/genitourinary system
L162.00	Unspecified renal disease in pregnancy
L162.11	Albuminuria in pregnancy without hypertension
L162.12	Nephropathy NOS in pregnancy without hypertension
L162.13	Uraemia in pregnancy without hypertension
L162000	Unspecified renal disease in pregnancy unspecified
L162100	Unspecified renal disease in pregnancy - delivered
L165.00	Asymptomatic bacteriuria in pregnancy
L165200	Asymptomatic bacteriuria in pregnancy – del with p/n comp
L165300	Asymptomatic bacteriuria in pregnancy – not delivered

L165z00	Asymptomatic bacteriuria in pregnancy NOS
L166.00	Genitourinary tract infections in pregnancy
L166.11	Cystitis of pregnancy
L166000	Genitourinary tract infection in pregnancy unspecified
L166100	Genitourinary tract infection in pregnancy - delivered
L166300	Genitourinary tract infection in pregnancy – not delivered
L166400	Genitourinary tract infection in pregnancy with p/n comp
L166500	Infections of kidney in pregnancy
L166600	Urinary tract infection following delivery
L166700	Infections of the genital tract in pregnancy
L166800	Urinary tract infection complicating pregnancy
L166z00	Genitourinary tract infection in pregnancy NOS
L166z11	UTI - urinary tract infection in pregnancy
L16A.00	Glycosuria during pregnancy
L16A000	Glycosuria during pregnancy unspecified
L16A100	Glycosuria during pregnancy - delivered
L16A200	Glycosuria during pregnancy - delivered with p/n comp
L16A300	Glycosuria during pregnancy - not delivered
L16Az00	Glycosuria during pregnancy NOS
L177.00	Infections of bladder in pregnancy
L178.00	Infections of urethra in pregnancy
L1y..00	Complications of pregnancy/childbirth/puerperium OS
L1z..00	Complications of pregnancy/childbirth/puerperium NOS
L3z..00	Complications of labour and delivery NOS
PDz..00	Urinary system anomalies NOS
PDz0.00	Unspecified anomaly of kidney
PDz2.00	Unspecified anomaly of bladder
PDz3.00	Unspecified anomaly of urethra
SP07Q00	Catheter-associated urinary tract infection
SP07Q11	CAUTI - catheter-associated urinary tract infection

Table A 7: Read codes for skin conditions.

Read code	Read term
14F..00	H/O: skin disorder
14F3.00	H/O: chronic skin ulcer
14F4.00	H/O: Admission in last year for diabetes foot problem
14F5.00	H/O: venous leg ulcer
14F6.00	H/O: foot ulcer
14F7.00	H/O: arterial lower limb ulcer
14FZ.00	H/O: skin disease NOS
1D14.00	C/O: a rash
1N0..00	Skin symptoms
1N00.00	Change in skin lesion
1N04.00	Itching of skin lesion
1N05.00	Mottling of skin
2F7..00	O/E - pustules
2F72.00	O/E - pustules present
2F73.00	O/E - purulent pustules
2F74.00	O/E - deep seated pustules
2F75.00	O/E - follicular pustules
2F7Z.00	O/E - pustules NOS
2FD..00	O/E - skin cyst
2FD2.00	O/E - skin cyst present
2FD2000	O/E - eyebrow cyst present
2FD2100	O/E - scalp cyst present
2FDZ.00	O/E - skin cyst NOS
2FF..00	O/E - skin ulcer
2FF2.00	O/E - skin ulcer present
2FF3.00	O/E - depth of ulcer
2FFZ.00	O/E - skin ulcer NOS
2G2A.00	Tinel's sign
2G35.00	O/E - nails - pitting
2G37.00	Splitting toenail
2G48.00	O/E - ankle ulcer
2G51000	Foot abnormality - diabetes related
2G54.00	O/E - Right foot ulcer
2G55.00	O/E - Left foot ulcer
2G5A.00	O/E - Right diabetic foot at risk
2G5B.00	O/E - Left diabetic foot at risk
2G5C.00	Foot abnormality - diabetes related
2G5d.00	O/E - Left diabetic foot at increased risk
2G5E.00	O/E - Right diabetic foot at low risk
2G5e.00	O/E - Right diabetic foot at increased risk
2G5F.00	O/E - Right diabetic foot at moderate risk
2G5G.00	O/E - Right diabetic foot at high risk
2G5H.00	O/E - Right diabetic foot - ulcerated

2G5I.00	O/E - Left diabetic foot at low risk
2G5J.00	O/E - Left diabetic foot at moderate risk
2G5K.00	O/E - Left diabetic foot at high risk
2G5L.00	O/E - Left diabetic foot - ulcerated
2G5S.00	O/E - right healed foot ulcer
2G5V.00	O/E - right chronic diabetic foot ulcer
2G5W.00	O/E - left chronic diabetic foot ulcer
2G64.00	O/E - infected toe
A35..00	Erysipelas
A90..00	Congenital syphilis
A900.00	Early congenital syphilis with symptoms
A900.12	Congenital syphilitic choroiditis
A900.13	Congenital syphilitic chronic coryza
A900.14	Congenital syphilitic epiphysitis
A900.16	Congenital syphilitic osteochondritis
A901.00	Early latent congenital syphilis
A902.00	Early congenital syphilis NOS
A903.00	Syphilitic interstitial keratitis
A904.00	Juvenile neurosyphilis
A904200	Congenital syphilitic meningitis
A905.00	Other late congenital syphilis
A905000	Congenital syphilitic gumma
A905100	Hutchinson's teeth
A905200	Syphilitic saddle nose
A905300	Late congenital syphilitic oculopathy
A906.00	Latent late congenital syphilis
A907.00	Unspecified late congenital syphilis
A90z.00	Congenital syphilis NOS
A913.00	Secondary syphilis of skin or mucus membranes
A913000	Secondary syphilis of anus
A913300	Secondary syphilis of skin
A913400	Secondary syphilis of tonsils
A913500	Secondary syphilis of vulva
A913z00	Secondary syphilis of skin or mucus membranes NOS
A92..00	Latent early syphilis
A92z.00	Latent early syphilis NOS
M0...00	Skin and subcutaneous tissue infections
M00..00	Carbuncle
M000.00	Carbuncle of face
M000000	Carbuncle of ear
M000100	Carbuncle of face (excluding eye)
M000200	Carbuncle of nasal septum
M000300	Carbuncle of temple region
M000z00	Carbuncle of face NOS
M001.00	Carbuncle of neck
M002.00	Carbuncle of trunk
M002000	Carbuncle of chest wall
M002100	Carbuncle of breast

M002200	Carbuncle of back
M002300	Carbuncle of abdominal wall
M002400	Carbuncle of umbilicus
M002500	Carbuncle of flank
M002600	Carbuncle of groin
M002700	Carbuncle of perineum
M002z00	Carbuncle of trunk NOS
M003.00	Carbuncle of upper arm and forearm
M003000	Carbuncle of shoulder
M003100	Carbuncle of axilla
M003200	Carbuncle of upper arm
M003300	Carbuncle of elbow
M003400	Carbuncle of forearm
M003z00	Carbuncle of upper arm and forearm NOS
M004.00	Carbuncle of hand
M004000	Carbuncle of wrist
M004100	Carbuncle of thumb
M004200	Carbuncle of finger
M004z00	Carbuncle of hand NOS
M005.00	Carbuncle of buttock
M005000	Carbuncle of anus
M005100	Carbuncle of gluteal region
M005z00	Carbuncle of buttock NOS
M006.00	Carbuncle of leg (excluding foot)
M006000	Carbuncle of hip
M006100	Carbuncle of thigh
M006200	Carbuncle of knee
M006300	Carbuncle of lower leg
M006400	Carbuncle of ankle
M006z00	Carbuncle of leg (excluding foot) NOS
M007.00	Carbuncle of foot
M007100	Carbuncle of heel
M007200	Carbuncle of toe
M007z00	Carbuncle of foot NOS
M00y.00	Carbuncle of other specified site
M00y000	Carbuncle of head (excluding face)
M00yz00	Carbuncle of other specified site NOS
M00z.00	Carbuncle NOS
M01..00	Furuncle - boil
M010.00	Boil of face
M010000	Boil of ear
M010100	Boil of face (excluding eye)
M010200	Boil of nasal septum
M010300	Boil of temple region
M010400	Boil of external nose
M010z00	Boil of face NOS
M011.00	Boil of neck
M012.00	Boil of trunk

M012000	Boil of chest wall
M012100	Boil of breast
M012200	Boil of back
M012300	Boil of abdominal wall
M012400	Boil of umbilicus
M012500	Boil of flank
M012600	Boil of groin
M012700	Boil of perineum
M012z00	Boil of trunk NOS
M013.00	Boil of upper arm and forearm
M013000	Boil of shoulder
M013100	Boil of axilla
M013200	Boil of upper arm
M013300	Boil of elbow
M013400	Boil of forearm
M013z00	Boil of upper arm and forearm NOS
M014.00	Boil of hand
M014000	Boil of wrist
M014100	Boil of thumb
M014200	Boil of finger
M014z00	Boil of hand NOS
M015.00	Boil of buttock
M015000	Boil of anus
M015100	Boil of gluteal region
M015z00	Boil of buttock NOS
M016.00	Boil of leg (excluding foot)
M016000	Boil of hip
M016100	Boil of thigh
M016200	Boil of knee
M016300	Boil of lower leg
M016400	Boil of ankle
M016z00	Boil of leg (excluding foot) NOS
M017.00	Boil of foot
M017000	Boil of foot unspecified
M017100	Boil of heel
M017200	Boil of toe
M017z00	Boil of foot NOS
M01y.00	Boil of other specified site
M01y000	Boil of head (excluding face)
M01yz00	Boil of other specified site NOS
M01z.00	Boil NOS
M01z.11	Recurrent boils
M01z.12	Boils of multiple sites
M01z000	Multiple boils
M02..00	Cellulitis and abscess of finger and toe
M020.00	Cellulitis and abscess of finger
M020000	Cellulitis and abscess of finger unspecified
M020100	Finger pulp abscess

M020111	Felon
M020112	Whitlow
M020200	Onychia of finger
M020300	Paronychia of finger
M020311	Perionychia of finger
M020400	Finger web space infection
M020500	Pulp space infection of finger/thumb
M020z00	Cellulitis and abscess of finger NOS
M021.00	Cellulitis and abscess of toe
M021000	Cellulitis and abscess of toe unspecified
M021100	Onychia of toe
M021200	Paronychia of toe
M021300	Pulp space infection of toe
M021z00	Cellulitis and abscess of toe NOS
M021z11	Perionychia of toe
M02z.00	Cellulitis and abscess of digit NOS
M02z.11	Nail infection NOS
M02z.12	Paronychia
M02z.13	Infected nailfold
M02z.14	Nailfold infected
M03..00	Other cellulitis and abscess
M03..11	Abscess of skin area excluding digits of hand or foot
M03..12	Acute lymphangitis of skin excluding digits of hand or foot
M03..13	Cellulitis of skin area excluding digits of hand or foot
M030.00	Cellulitis and abscess of face
M030000	Cellulitis and abscess of cheek (external)
M030011	Cellulitis and abscess of cheek
M030100	Cellulitis and abscess of nose (external)
M030111	Cellulitis and abscess of nose
M030200	Cellulitis and abscess of chin
M030300	Cellulitis and abscess of submandibular region
M030400	Cellulitis and abscess of forehead
M030500	Cellulitis and abscess of temple region
M030600	Cellulitis of face
M030z00	Cellulitis and abscess of face NOS
M031.00	Cellulitis and abscess of neck
M032.00	Cellulitis and abscess of trunk
M032000	Cellulitis and abscess of chest wall
M032100	Cellulitis and abscess of breast
M032200	Cellulitis and abscess of back
M032300	Cellulitis and abscess of abdominal wall
M032400	Cellulitis and abscess of umbilicus
M032500	Cellulitis and abscess of flank
M032600	Cellulitis and abscess of groin
M032700	Cellulitis and abscess of perineum
M032800	Cellulitis of trunk
M032z00	Cellulitis and abscess of trunk NOS
M033.00	Cellulitis and abscess of arm

M033000	Cellulitis and abscess of shoulder
M033100	Cellulitis and abscess of axilla
M033200	Cellulitis and abscess of upper arm
M033300	Cellulitis and abscess of elbow
M033400	Cellulitis and abscess of forearm
M033z00	Cellulitis and abscess of arm NOS
M034.00	Cellulitis and abscess of hand excluding digits
M034.11	Cellulitis and abscess of hand
M034000	Cellulitis and abscess of hand unspecified
M034011	Abscess of dorsum of hand
M034012	Abscess of palm of hand
M034013	Cellulitis of dorsum of hand
M034014	Cellulitis of palm of hand
M034100	Cellulitis and abscess of wrist
M034z00	Cellulitis and abscess of hand NOS
M035.00	Cellulitis and abscess of buttock
M036.00	Cellulitis and abscess of leg excluding foot
M036.11	Cellulitis and abscess of leg
M036000	Cellulitis and abscess of hip
M036100	Cellulitis and abscess of thigh
M036200	Cellulitis and abscess of knee
M036300	Cellulitis and abscess of lower leg
M036400	Cellulitis and abscess of ankle
M036z00	Cellulitis and abscess of leg NOS
M037.00	Cellulitis and abscess of foot excluding toe
M037.11	Cellulitis and abscess of foot
M037000	Cellulitis and abscess of foot unspecified
M037100	Cellulitis and abscess of heel
M037200	Cellulitis in diabetic foot
M037z00	Cellulitis and abscess of foot NOS
M038.00	Cellulitis of external ear
M03y.00	Other specified cellulitis and abscess
M03y000	Cellulitis and abscess of head unspecified
M03y011	Abscess of scalp
M03z.00	Cellulitis and abscess NOS
M03z000	Cellulitis NOS
M03z100	Abscess NOS
M03zz00	Cellulitis and abscess NOS
M03zz11	Acute lymphangitis NOS
M04..00	Acute lymphadenitis
M04..11	Acute abscess lymph node
M04..12	Acute adenitis
M040.00	Acute lymphadenitis of trunk
M041.00	Acute lymphadenitis of upper limb
M042.00	Acute lymphadenitis of lower limb
M043.00	Acute lymphadenitis of face, head and neck
M05..00	Impetigo
M050.00	Impetigo contagiosa unspecified

M051.00	Impetigo contagiosa bullosa
M052.00	Impetigo contagiosa gyrata
M053.00	Impetigo circinata
M054.00	Impetigo neonatorum
M055.00	Impetigo simplex
M056.00	Impetigo follicularis
M057.00	Chronic symmetrical impetigo
M05z.00	Impetigo NOS
M06..00	Pilonidal sinus/cyst
M060.00	Pilonidal cyst with abscess
M061.00	Pilonidal cyst with no abscess
M061.11	Dermal sinus
M062.00	Pilonidal sinus with abscess
M063.00	Pilonidal sinus without abscess
M06z.00	Pilonidal sinus/cyst NOS
M07..00	Other local infections of skin and subcutaneous tissue
M070.00	Pyoderma
M070.11	Purulent dermatitis
M070000	Pyoderma chancriforme
M070100	Pyoderma faciale
M070200	Pyoderma gangrenosum
M070300	Pyoderma ulcerosum tropicalum
M070z00	Pyoderma NOS
M071.00	Pyogenic granuloma
M071000	Pyogenic granuloma unspecified
M071100	Pyogenic progressive granuloma
M071200	Granuloma telangiectaticum
M071300	Umbilical granuloma
M071z00	Pyogenic granuloma NOS
M072.00	Erythrasma
M073.00	Scalp infection
M07y.00	Local infection of skin or subcutaneous tissue OS
M07y.11	Pustular eczema
M07y000	Pustular bacterid
M07y100	Ecthyma
M07y200	Dermatitis vegetans
M07y300	Perleche
M07y400	Pitted keratolysis
M07y500	Inflammation of scar
M07yz00	Other spec local skin/subc infection NOS
M07yz11	Infection toe
M07yz12	Infection foot
M07yz13	Infection finger
M07z.00	Local infection skin/subcut tissue NOS
M07z.11	Infected insect bite
M07z.12	Infected skin ulcer
M07z.13	Septic spots
M07z.14	Infected dermatitis

M07z.15	Sinus
M07z000	Infection foot
M07z100	Infection toe
M07z200	Infection finger
M08..00	Cutaneous cellulitis
M080.00	[X]Cellulitis of finger and toe
M080.11	[X]Nail bed infection
M080.12	[X]Septic thumb
M080.13	[X]Cellulitis of thumb
M081.00	[X]Cellulitis of other parts of limb
M082.00	Cellulitis of face
M083.00	Cellulitis of trunk
M084.00	[X]Cellulitis of breast
M085.00	Cellulitis of leg
M086.00	Cellulitis of ankle
M087.00	Chronic paronychia
M088.00	Cellulitis of arm
M089.00	Cellulitis of neck
M08A.00	Cellulitis of axilla
M08B.00	Cellulitis of foot
M08C.00	Cellulitis of toe
M08y.00	[X]Cellulitis of other sites
M09..00	Cutaneous abscess
M090.00	[X]Abscess of face
M091.00	[X]Abscess of neck
M092.00	[X]Abscess of trunk
M092000	[X]Abscess of buttock
M092100	[X]Abdominal wall abscess
M092200	[X]Perineal abscess
M093.00	[X]Abscess of buttock
M094.00	[X]Abscess of limb
M094000	[X]Abscess of axilla
M095.00	Skin abscess
M09y.00	[X]Abscess of other site
M0y..00	Other specified infections of skin or subcutaneous tissue
M0z..00	Skin and subcut tissue infection NOS
M0z..11	Infected sebaceous cyst
M111.00	Atopic dermatitis/eczema
M153.00	Rosacea
M153000	Acne rosacea
M153100	Rhinophyma
M153200	Rosacea hypertrophica
M153300	Lupoid rosacea
M153400	Ocular rosacea
M153500	Perioral dermatitis
M153511	Circumoral dermatitis
M153600	Periocular dermatitis
M153z00	Rosacea NOS

M244.00	Folliculitis
M25y100	Hidradenitis
M25y111	Hidradenitis suppurativa
M25y600	Acne keloid
M25yX00	Apocrine sweat disorder, unspecified
M26..00	Sebaceous gland diseases
M260.00	Acne varioliformis
M260000	Acne frontalis
M260z00	Acne varioliformis NOS
M260z11	Acne necrotica
M261.00	Other acne
M261000	Acne vulgaris
M261011	Blackhead
M261012	Comedo
M261100	Acne conglobata
M261200	Bromine acne
M261300	Chlorine acne
M261400	Iodine acne
M261500	Colloid acne
M261600	Cystic acne
M261700	Acne neonatorum
M261800	Infantile acne
M261900	Occupational acne
M261A00	Pustular acne
M261B00	Steroid acne
M261C00	Tropical acne
M261D00	Acne urticata
M261E00	Acne excoriee des jeunes filles
M261F00	Acne fulminans
M261G00	Acne agminata
M261H00	Acne keloid
M261J00	Acne necrotica
M261K00	Acne keloidalis
M261X00	Acne, unspecified
M261z00	Other acne NOS
M262.00	Sebaceous cyst - wen
M262.11	Keratin cyst
M262.12	Sebaceous cyst
M262000	Trichilemmal cyst
M262100	Pilar cyst
M262200	Pilar cyst of scalp
M262211	Sebaceous cyst of scalp
M263.00	Seborrhoea
M263000	Seborrhoea corporis
M263100	Seborrhoea faciei
M263200	Seborrhoea nasi
M263300	Seborrhoea oleosa
M263400	Post-encephalitic seborrhoea

M263z00	Seborrhoea NOS
M26y.00	Other specified sebaceous gland diseases
M26y000	Asteatosis cutis
M26y200	Giant comedo
M26y300	Fordyce spots
M26y400	Sebaceous gland hypertrophy
M26yz00	Other sebaceous gland diseases NOS
M26z.00	Sebaceous gland diseases NOS
M2yz.11	Skin lesion
Myu6800	[X]Other acne
R02..00	[D]Symptoms affecting skin and other integumentary tissue
R021.00	[D]Rash and other nonspecific skin eruption
R021000	[D]Exanthem
R021100	[D]Rash on genitals
R021z00	[D]Rash and other nonspecific skin eruption NOS
R021z11	[D]Spots
R022.00	[D]Local superficial swelling, mass or lump
R022000	[D]Swelling, local and superficial
R022100	[D]Mass, localized and superficial
R022200	[D]Lump, localized and superficial
R022300	[D]Nodule, subcutaneous
R022400	[D]Localized swelling, mass and lump, upper limb
R022500	[D]Localized swelling, mass and lump, lower limb
R022600	[D]Localized swelling, mass and lump, multiple sites
R022700	[D]Axillary lump
R022800	[D]Lump on back
R022900	[D]Foot lump
R022A00	[D]Shoulder lump
R022B00	[D]Lump on hand
R022C00	[D]Lump on knee
R022D00	[D]Lump on leg
R022E00	[D]Lump on shin
R022F00	[D]Lump on thigh
R022G00	[D]Finger lump
R022H00	[D]Wrist lump
R022I00	[D]Toe lump
R022J00	[D]Subungual swelling
R022K00	[D]Buttock swelling
R022z00	[D]Local superficial swelling, mass or lump NOS
R02z.00	[D]Skin symptoms NOS

Table A 8: Read codes for eye conditions.

Read code	Read term
1486	H/O: iritis
148Z.00	H/O: eye disorder NOS
F400.00	Purulent endophthalmitis
F400100	Acute endophthalmitis
F400200	Panophthalmitis
F400300	Chronic endophthalmitis
F400400	Vitreous abscess
F400500	Eye infection
F400z00	Purulent endophthalmitis NOS
F44..00	Disorders of iris and ciliary body
F44..11	Ciliary body disorders
F44..12	Iridocyclitis
F440.00	Acute and subacute iridocyclitis
F440.11	Iritis - acute
F440000	Unspecified acute iridocyclitis
F440100	Unspecified subacute iridocyclitis
F440200	Primary iridocyclitis
F440300	Recurrent iridocyclitis
F440400	Secondary infected iridocyclitis
F440600	Hypopyon
F440700	Diabetic iritis
F440z00	Acute or subacute iritis NOS
F441.00	Chronic iridocyclitis
F441.11	Chronic iritis
F441000	Unspecified chronic iridocyclitis
F441100	Chronic iridocyclitis due to disease EC
F441200	Chronic anterior uveitis
F441z00	Chronic iridocyclitis NOS
F442.00	Certain types of iridocyclitis
F442000	Fuchs' heterochromic cyclitis
F442200	Lens-induced iridocyclitis
F442300	Vogt-Koyanagi syndrome
F442z00	Certain types of cyclitis NOS
F443.00	Unspecified iridocyclitis
F443.11	Uveitis NOS
F443000	Anterior uveitis
F443100	Iritis
F444.00	Iris and ciliary body vascular disorders
F444z00	Iris and ciliary body vascular disorders NOS
F446.11	Uveal cysts
F446000	Idiopathic cyst of iris, ciliary body or anterior chamber
F44y.00	Other iris and ciliary body disorders
F44yz00	Other iris or ciliary body disorder NOS
F44z.00	Iris or ciliary body disorder NOS
F48..00	Visual disturbances

F481.00	Subjective visual disturbances
F481000	Unspecified subjective visual disturbance
F481100	Sudden visual loss
F481400	Other transient visual loss
F481700	Photophobia
F481800	Other visual discomfort
F481C00	Photopsia
F481D00	Visual halos
F481E00	Refractive diplopia
F481F00	Refractive polyopia
F481G00	Other visual distortion
F481K00	Visual hallucinations
F482.00	Diplopia (double vision)
F483.00	Other binocular vision disorders
F483000	Unspecified binocular vision disorder
F484.00	Visual field defects
F484000	Unspecified visual field defect
F484z00	Visual field defects NOS
F485.00	Colour vision deficiency
F48y.00	Other specified visual disturbance
F48y000	Blurred vision NOS
F48y011	Cloudy vision NOS
F48y012	Dull vision NOS
F48yz00	Other specified visual disturbance NOS
F48z.00	Visual disturbance NOS
F4A..00	Keratitis
F4A..11	Keratoconjunctivitis
F4A0.00	Corneal ulcer
F4A0.11	Dendritic ulcer
F4A0000	Unspecified corneal ulcer
F4A0100	Marginal corneal ulcer
F4A0200	Ring corneal ulcer
F4A0300	Central corneal ulcer
F4A0400	Hypopyon ulcer
F4A0411	Serpiginous ulcer
F4A0600	Perforated corneal ulcer
F4A0700	Mooren's ulcer
F4A0z00	Corneal ulcer NOS
F4A1.00	Dendritic keratitis
F4A2.00	Other superficial keratitis without conjunctivitis
F4A2000	Unspecified superficial keratitis
F4A2100	Punctate keratitis
F4A2111	Keratitic precipitates
F4A2112	Thygeson's superficial punctate keratitis
F4A2200	Nummular keratitis
F4A2300	Striate keratitis
F4A2400	Macular keratitis NOS
F4A2500	Filamentary keratitis

F4A2711	Arc-welders' keratitis
F4A2800	Photokeratitis NOS
F4A2z00	Other superficial keratitis without conjunctivitis NOS
F4A3.00	Specific keratoconjunctivitis
F4A3000	Phlyctenular keratoconjunctivitis
F4A3100	Vernal conjunctivitis of limbus and cornea
F4A3200	Keratoconjunctivitis sicca (excluding Sjogren's syndrome)
F4A3300	Exposure keratoconjunctivitis
F4A3400	Neurotrophic keratoconjunctivitis
F4A3z00	Specific keratoconjunctivitis NOS
F4A4.00	Other keratoconjunctivitis
F4A4000	Unspecified keratoconjunctivitis
F4A4100	Keratitis or keratoconjunctivitis in other exanthemata
F4A4z00	Other keratoconjunctivitis NOS
F4A5.00	Interstitial and deep keratitis
F4A5000	Unspecified interstitial keratitis
F4A5100	Diffuse interstitial keratitis
F4A5300	Corneal abscess
F4A5400	Keratitis due to syphilis
F4A5500	Keratitis due to tuberculosis
F4A5z00	Interstitial and deep keratitis NOS
F4Ay.00	Other forms of keratitis
F4Az.00	Keratitis NOS
F4B..00	Corneal opacity and other disorders of cornea
F4B..11	Corneal disorders
F4B0.00	Corneal scars and opacities
F4B0z00	Corneal scar or opacity NOS
F4B1.00	Corneal pigmentations and deposits
F4B1z00	Corneal pigmentation or deposit NOS
F4B2.00	Corneal oedema
F4B2z00	Corneal oedema NOS
F4B3.00	Corneal membrane changes
F4B3z00	Corneal membrane changes NOS
F4B4.00	Corneal degenerations
F4B4z00	Corneal degenerations NOS
F4B5000	Corneal dystrophy unspecified
F4B7100	Corneal ectasia
F4B7300	Corneal staphyloma
F4By000	Corneal hypoaesthesia
F4By100	Corneal anaesthesia
F4Bz.00	Corneal disorder NOS
F4C0.00	Acute conjunctivitis
F4C0.11	Eye infection
F4C0.12	Conjunctivitis
F4C0000	Unspecified acute conjunctivitis
F4C0011	Conjunctivitis
F4C0100	Serous conjunctivitis
F4C0200	Acute follicular conjunctivitis

F4C0300	Acute mucopurulent conjunctivitis
F4C0311	Sticky eye
F4C0400	Catarrhal conjunctivitis
F4C0500	Pseudomembranous conjunctivitis
F4C0511	Membranous conjunctivitis
F4C0600	Acute atopic conjunctivitis
F4C0611	Acute allergic conjunctivitis
F4C0z00	Acute conjunctivitis NOS
F4C1.00	Chronic conjunctivitis
F4C1000	Unspecified chronic conjunctivitis
F4C1100	Simple chronic conjunctivitis
F4C1200	Chronic follicular conjunctivitis
F4C1300	Vernal conjunctivitis
F4C1z00	Chronic conjunctivitis NOS
F4C2.00	Blepharoconjunctivitis
F4C2000	Unspecified blepharoconjunctivitis
F4C2100	Angular blepharoconjunctivitis
F4C2200	Contact blepharoconjunctivitis
F4C2z00	Blepharoconjunctivitis NOS
F4C3.00	Other and unspecified conjunctivitis
F4C3000	Unspecified conjunctivitis
F4C3100	Rosacea conjunctivitis
F4C3200	Conjunctivitis with mucocutaneous disorder
F4C3300	Bacterial conjunctivitis
F4C3z00	Other conjunctivitis NOS
F4Cy.00	Other conjunctival disorders
F4Cy000	Filarial infection of conjunctiva
F4Cy100	Ocular pemphigoid
F4D..00	Inflammation of eyelids
F4D0.00	Blepharitis
F4D0.11	Cellulitis of eyelids
F4D0000	Unspecified blepharitis
F4D0100	Ulcerative blepharitis
F4D0200	Squamous blepharitis
F4D0z00	Blepharitis NOS
F4D1.00	Hordeolum and other deep inflammation of eyelid
F4D1000	Hordeolum externum (stye)
F4D1100	Hordeolum internum (infected meibomian cyst)
F4D1111	Meibomian cyst infected
F4D1200	Abscess of eyelid
F4D1211	Boil of eyelid
F4D1212	Furuncle of eyelid
F4D1300	Meibomianitis
F4D1400	Cellulitis of eyelid
F4D1z00	Hordeolum and other deep inflammation of eyelid NOS
F4D2.00	Chalazion (meibomian cyst)
F4D3000	Eczematous eyelid dermatitis
F4D3100	Contact or allergic eyelid dermatitis

F4D3111	Allergic dermatitis - eyelid
F4D3112	Contact eczema - eyelids
F4D4.00	Infective eyelid dermatitis of types resulting in deformity
F4D5.00	Other eyelid infective dermatitis
F4D6.00	Parasitic eyelid infestation
F4Dy.00	Other eyelid inflammation
F4Dy000	Ulcer of eyelid
F4Dz.00	Eyelid inflammation NOS
F4F0.00	Dacryoadenitis
F4F0000	Unspecified dacryoadenitis
F4F0100	Acute dacryoadenitis
F4F0200	Chronic dacryoadenitis
F4F0z00	Dacryoadenitis NOS
F4F3.00	Acute and unspecified inflammation of lacrimal passages
F4F3.11	Dacryocystitis - acute
F4F3000	Unspecified dacryocystitis
F4F3100	Acute lacrimal canaliculitis
F4F3200	Acute dacryocystitis
F4F3300	Phlegmonous dacryocystitis
F4F3z00	Dacryocystitis NOS
F4F6000	Lacrimal fistula
F4G0.00	Acute inflammation of orbit
F4G0000	Unspecified acute orbit inflammation
F4G0100	Orbital cellulitis
F4G0200	Orbital abscess
F4G0300	Orbital periostitis
F4G0400	Orbital osteomyelitis
F4G0500	Tenonitis
F4G0z00	Acute inflammation of orbit NOS
F4G1.00	Chronic inflammation of orbit
F4G1000	Unspecified chronic inflammation of orbit
F4G1300	Parasitic infestation of orbit
F4G1z00	Chronic inflammation of orbit NOS
F4K0.00	Scleritis and episcleritis
F4K0.11	Episcleritis
F4K0.12	Scleritis
F4K0000	Unspecified scleritis
F4K0100	Episcleritis periodica fugax
F4K0200	Episcleritis periodica fugax
F4K0300	Anterior scleritis
F4K0400	Scleromalacia perforans
F4K0500	Sclerokeratitis
F4K0600	Brawny scleritis
F4K0700	Posterior scleritis
F4K0711	Sclerotenonitis
F4K0800	Scleral abscess
F4K0z00	Scleritis or episcleritis NOS
F4Kz.00	Eye and adnexa disorder NOS

F4Kz000	Unspecified disorder of eye
F4Kz100	Eye pain NOS
F4Kz200	Swelling of eye NOS
F4Kz300	Mass of eye NOS
F4Kz400	Redness of eye NOS
F4Kz411	Red eye NOS
F4Kz500	Discharge of eye NOS
F4Kzz00	Ill-defined eye disorder NOS
SD81000	Corneal abrasion
SG00.00	Corneal foreign body
SH0..12	Corneal burns

Appendix D: Searching strategies for systematic review

Database: Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present (last update 20/01/2020)

#1. Lower Respiratory Infections (including pneumonia)

1. (RTI or rti or (respiratory tract\$ adj3 infect\$)).ab,ti. (24,121)
2. ((chest\$ or thorac\$) adj5 infect\$).ab,ti. (3,911)
3. (bronchi\$ or trachei\$).ab,ti. (130,796)
4. ((pulmonary or lung) adj3 (inflammation\$ or infect\$)).ab,ti. (37,243)
5. exp Pneumonia/ (89,835)
6. (LRTI or (low\$ respiratory tract\$ adj3 infect\$)).ab,ti. (7,028)
7. (cough\$ or bronchit\$).ab,ti. (67,657)
8. 1 or 2 or 3 or 4 or 5 or 6 or 7 (302,314)

#2. Risk factor

9. (risk factor\$ or ((patient\$ or population) adj3 risk)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (1,217,264)

#3. Prediction Model

10. Models, Theoretical/ or Linear Models/ or Models, Biological/ or Logistic Models/ or Models, Molecular/ (913,060)
11. (predic\$ or prognos\$ or probabilit\$ or valid\$).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (2,949,035)
12. 9 or 10 or 11 (4,464,682)
13. 8 and 12 (51,004)

#4. Primary care setting

14. (general practice\$ or GP or GP's or GP surger\$ or family practice\$ or out patient\$ or out-patient\$ or community or (community adj3 setting\$) or ambulatory).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (829,471)
15. (general practitioner\$ or family physician\$ or primary care physician\$ or general physician\$ or family doctor\$).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (87,583)
16. ((primary adj3 \$care) or (general adj3 \$care) or (community adj3 \$care)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (186,460)
17. 14 or 15 or 16 (969,078)
18. 13 and 17 (6,515)
19. limit 18 to (yr="2010 -Current" and english) (2,944)

#1. Lower Respiratory Infections (including pneumonia)

1. (RTI or rti or (respiratory tract\$ adj3 infect\$)).ab,ti. (36,617)
2. ((chest\$ or thorac\$) adj5 infect\$).ab,ti. (6,589)
3. (bronchi\$ or trachei\$).ab,ti. (204,769)
4. ((pulmonary or lung) adj3 (inflammation\$ or infect\$)).ab,ti. (57,875)
5. exp Pneumonia/ (311,302)
6. (LRTI or (low\$ respiratory tract\$ adj3 infect\$)).ab,ti. (10,372)
7. (cough\$ or bronchit\$).ab,ti. (116,372)
8. 1 or 2 or 3 or 4 or 5 or 6 or 7 (617,325)

#2. Risk factor

9. (risk factor\$ or ((patient\$ or population) adj3 risk)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (1,692,686)

#3. Prediction Model

10. Models, Theoretical/ or Linear Models/ or Models, Biological/ or Logistic Models/ or Models, Molecular/ (420,442)
11. (predic\$ or prognos\$ or probabilit\$ or valid\$).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (4,003,125)
12. 9 or 10 or 11 (5,530,594)
13. 8 and 12 (116,947)

#4. Primary care setting

14. (general practice\$ or GP or GP's or GP surger\$ or family practice\$ or out patient\$ or out-patient\$ or community or (community adj3 setting\$) or ambulatory).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (1,042,779)

15. (general practitioner\$ or family physician\$ or primary care physician\$ or general physician\$ or family doctor\$).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (158,013)
16. ((primary adj3 \$care) or (general adj3 \$care) or (community adj3 \$care)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (317,738)
17. 14 or 15 or 16 (1,283,242)
18. 13 and 17 (13,073)
19. limit 18 to (yr="2010 -Current" and english) (8,637)

Appendix E: Excluded studies by full text

The full texts of following studies were reviewed for intended inclusion but failed to meet inclusion criteria due to reasons below: 1: Ineligible population (i.e. HIV patients); 2: Ineligible clinical setting (i.e. hospital setting or nursing home); 3: Ineligible study design (i.e. case report, biomedical experimental research, qualitative study or prediction model validation study); 4: Ineligible endpoint (i.e. hospital admission due to pneumonia); 5: Ineligible publication type (i.e. editorial letter); 6: Non-English language papers with English abstracts; 7: Observational studies with inadequate sample size for prediction modelling (i.e. event per variable is fewer than 10); 8: No full text available for access.

Table A 9: Excluded studies by full text

Study	Exclusion code	Note
1. (Pick et al., 2020) Pneumococcal serotype trends, surveillance and risk factors in UK adult pneumonia, 2013-18	4	Prediction endpoint being pneumococcal vaccination
2. (Nakajima et al., 2020) Association between oral candidiasis and bacterial pneumonia: A retrospective study	2	Nursing home residents
3. (Tashiro et al., 2019) Relationship between oral environment and development of pneumonia and acute viral respiratory infection in dependent older individuals	2	Nursing home residents
4. (Garcia Garrido et al., 2019) Incidence and Risk Factors for Invasive Pneumococcal Disease and Community-acquired Pneumonia in Human Immunodeficiency Virus-Infected Individuals in a High-income Setting	1	Study patient being HIV patients
5. (Clark et al., 2020) Plasma vitamin D, past chest illness, and risk of future chest illness in chronic spinal cord injury (SCI): a longitudinal observational study	4	chronic spinal cord injury (SCI) cohort; endpoint being chest illness
6. (Abelleira et al., 2019) Influenza A H1N1 Community-Acquired Pneumonia: Characteristics and Risk Factors-A Case-Control Study	2	Endpoint being influenza pneumonia out of hospitalized CAP patients
7. (Heath et al., 2019) Variable selection for early diagnosis of congenital heart disease using random forest entropy calculations	4	Endpoint being congenital heart failure
8. (Sungur Biteker et al., 2019) Right Heart Function in Community-Acquired Pneumonia	4	Endpoint being different phase of CAP

9. (Zhang et al., 2019) Utility of Blood Cultures in Pneumonia	2	Hospital setting
10. (Hadda et al., 2019) Severe community acquired pneumonia: Prediction of outcome	5	Editorial
11. (Gonzalez Del Castillo et al., 2019) Risk stratification of patients with pneumonia	8	No full text
12. (Ozlek et al., 2019) The risk stratification in community-acquired pneumonia	8	No full text
13. (Van Buynder, 2019) Reducing pneumococcal risk in people aged 65 years and over	8	Review
14. (Malek et al., 2019) Relationship between the serum level of C-reactive protein and severity and outcomes of community-acquired pneumonia	4	Endpoint being CAP severity and mortality
15. (Ham and Eun Song, 2019) A prospective study of presepsin as an indicator of the severity of community-acquired pneumonia in emergency departments: Comparison with pneumonia severity index and CURB-65 scores	4	Pneumonia severity prediction model comparison
16. (Akagi et al., 2019) Procalcitonin is not an independent predictor of 30-day mortality, albeit predicts pneumonia severity in patients with pneumonia acquired outside the hospital	4	Pneumonia severity prediction
17. (Tanzella et al., 2019) Optimal approaches to preventing severe community-acquired pneumonia	8	No full text
18. (Kalra et al., 2019) Elevated C-reactive protein increases diagnostic accuracy of algorithm-defined stroke-associated pneumonia in afebrile patients	4	Endpoint being stroke associated pneumonia
19. (Bastidas et al., 2019) CURB-65 validity through use of artificial intelligence for multiple outcomes in community acquired pneumonia	3	Model validity study
20. (Garin et al., 2019) Computed tomography scan contribution to the diagnosis of community-acquired pneumonia	2	Hospital setting
21. (Edelman et al., 2019) Association of Prescribed Opioids with Increased Risk of Community-Acquired Pneumonia among Patients with and Without HIV	2	Hospital setting

22. (Rillera and Manuel, 2019) Association of procalcitonin levels and risk of mortality in adults with hospital acquired pneumonia and high risk community acquired pneumonia	4	Endpoint being mortality due to CAP and HAP
23. (Karakioulaki and Stolz, 2019b) Biomarkers in pneumonia-beyond procalcitonin	5	Review
24. (Karakioulaki and Stolz, 2019a) Biomarkers and clinical scoring systems in community-acquired pneumonia	5	Review
25. (Cohen et al., 2019) Associations between Community-Acquired Pneumonia and Proton Pump Inhibitors in the Laryngeal/Voice-Disordered Population	1	Patient cohort being laryngeal/voice-disordered population
26. (Eizadi-Mood et al., 2018) Risk factors associated with aspiration pneumonia among the patients with drug intoxication	2	Hospital setting
27. (Lu et al., 2018) Link between community-acquired pneumonia and vitamin D levels in older patients	2	Hospital setting
28. (Siljan et al., 2018) Procalcitonin has wider applicability in community-acquired pneumonia than other biomarkers	4	End points being pneumonia severity and treatment outcome
29. (Lee and Song, 2018) Pneumococcal urinary antigen test use as a prognostic marker in patients admitted with community-acquired pneumonia: A propensity score matching study	4	Endpoint being CAP mortality
30. (Xu et al., 2017) eGFR and the Risk of Community-Acquired Infections	4	Composite endpoint
31. (Sadigov and Abdullayev, 2017) Corellation between emphysema and pneumonia risk in patients with chronic obstructive pulmonary disease	8	No full text
32. (Stern et al., 2017) Corticosteroids for pneumonia	4	Endpoint being treatment effect for pneumonia

33. (Cho et al., 2017) Prognostic significance of nutritional risk in the elderly patients with community acquired pneumonia	8	No full text
34. (Faverio and Sibila, 2017) New biomarkers in community-acquired pneumonia: Another step in improving outcome prediction	5	Editorial
35. (Al-Helou et al., 2016) Validation of the modified CRB-65 pneumonia severity index as a prognostic tool	2	Hospital setting
36. (Almirall et al., 2016) Risk Factors for Community-acquired Pneumonia in Adults: A Review	5	Review
37. (Stolz, 2016) Procalcitonin in Severe Community-Acquired Pneumonia: Some Precision Medicine Ready for Prime Time	5	Editorial
38. (Minnaard et al., 2017) The added value of C-reactive protein measurement in diagnosing pneumonia in primary care: A meta-analysis of individual patient data	4	Prognostic value of single biomarker (CRP)
39. (Choby and Hunter, 2015) Respiratory infections: community-acquired pneumonia	8	No full text
40. (Dang et al., 2015) Recurrent pneumonia: a review with focus on clinical epidemiology and modifiable risk factors in elderly patients	4	End point being recurrent pneumonia after incident pneumonia
41. (Ticinesi et al., 2016) An investigation of multimorbidity measures as risk factors for pneumonia in elderly frail patients admitted to hospital	2	Hospital setting
42. (Schierenberg et al., 2016) External validation of prediction models for pneumonia in primary care patients with lower respiratory tract infection: An individual patient data meta-analysis	3	External validation study of pneumonia prediction models
43. (Lin et al., 2016) Increased risk of community-acquired pneumonia in COPD patients with comorbid cardiovascular disease	2	Unclear clinical setting
44. (Janson et al., 2018) Identifying the associated risks of pneumonia in COPD patients: ARCTIC an observational study	4	Pneumonia managed in both primary and secondary care

45. (Chatterjee et al., 2016) Anticholinergic medication use and risk of pneumonia in elderly adults: A nested case-control study	4	Pneumonia managed in both primary and secondary care
46. (Li et al., 2015) Severe pneumonia in the elderly: A multivariate analysis of risk factors	2	Hospitalized pneumonia
47. (Ishifuji et al., 2015) Medications associated with the incidence of recurrent pneumonia in Japanese elderly population	4	Endpoint being recurrent pneumonia
48. (Williams et al., 2015) Co-morbidity and pneumonia risk in COPD patients: A population database analysis of primary care patients	5	Spoken sessions: COPD weighs heavy on the heart
49. (Nose et al., 2015) Antipsychotic drug exposure and risk of pneumonia: A systematic review and meta-analysis of observational studies	4	Endpoint being pneumonia
50. (Smith et al., 2014) Infections diseases and stroke a novel point-of-care clinical risk score for predicting pneumonia in acute stroke care: a UK multicenter cohort study	5	Conference abstract
51. (Simonetti et al., 2014b) Impact of pre-hospital antibiotic use on community-acquired pneumonia	2	Hospitalized CAP patients
52. (Simonetti et al., 2014a) Management of community-acquired pneumonia in older adults	5	Review
53. (Sanz et al., 2014a) Does prolonged onset of symptoms have a prognostic significance in community-acquired pneumonia?	2	Hospitalized CAP patients
54. (Pick et al., 2014) Clinical characteristics of hospitalised patients misdiagnosed with community-acquired pneumonia	2	Hospitalized patients with initial CAP diagnosis
55. (Nemoto et al., 2014) Incremental prognostic predict ability of chest computed tomography in patients with community onset pneumonia	2	Hospitalized CAP patients
56. (Nie et al., 2014) Obesity survival paradox in pneumonia: A meta-analysis	4	All cause pneumonia as a composite endpoint

57. (Kew and Seniukovich, 2014) Inhaled steroids and risk of pneumonia for chronic obstructive pulmonary disease	4	All cause pneumonia as a composite endpoint
58. (Samokhvalov et al., 2010) Alcohol consumption as a risk factor for pneumonia: A systematic review and meta-analysis	4	Endpoint being CAP morbidity and /or mortality
59. (Trifiro, 2011) Antipsychotic drug use and community-acquired pneumonia	5	Report
60. (Doshi et al., 2011) Anemia and community-acquired pneumococcal pneumonia	2	Hospitalized CAP patients
61. (Giuliano et al., 2012) Are proton pump inhibitors associated with the development of community-acquired pneumonia? A meta-analysis	5	No full text
62. (Yende et al., 2013) Epidemiology and Long-term Clinical and Biologic Risk Factors for Pneumonia in Community-Dwelling Older Americans: Analysis of Three Cohorts	4	Endpoint being hospitalized pneumonia
63. (Dublin et al., 2012) Angiotensin-converting enzyme inhibitor use and pneumonia risk in community-dwelling older adults: Results from a population-based case-control study	4	Endpoint being pneumonia from outpatient and inpatient settings
64. (Van Vugt et al., 2013) Diagnosing pneumonia in patients with acute cough: Clinical judgment compared to chest radiography	3	Comparative study
65. (Cilloniz et al., 2013) Impact of age and comorbidity on cause and outcome in community-acquired pneumonia	4	Endpoint being causal pathogens and treatment outcomes of CAP
66. (Watanabe Tejada et al., 2013) Effect of comorbidities on clinical outcomes in low-risk curbed-65 patients	2	Inpatient CAP cohort
67. (Tanday, 2013) C-reactive protein could predict pneumonia in COPD	5	News
68. (Morillo et al., 2013) Computer-aided diagnosis of pneumonia in patients with chronic obstructive pulmonary disease	2	Hospitalized CAP patients

69. (Almirall et al., 2013) Relationship between the Use of Inhaled Steroids for Chronic Respiratory Diseases and Early Outcomes in Community-Acquired Pneumonia	4	Endpoint being CAP hospitalization
70. (Torres et al., 2013) Risk factors for community-acquired pneumonia in adults in Europe: A literature review	5	Review
71. (Eurich et al., 2013) Inhaled corticosteroids and risk of recurrent pneumonia: A population-based, nested case-control study	2	Emergency and inpatient departments
72. (Polverino et al., 2013) Influence of comorbidities on pneumococcal community-acquired pneumonia	4	Endpoints being subtypes of CAP
73. (Schubert et al., 2013) Observational study on diagnostics and treatment in community acquired pneumonia (CAP)	2	Hospitalized CAP
74. (Subramanian et al., 2013) Performance of SOAR (systolic blood pressure, oxygenation, age and respiratory rate) scoring criteria in community-acquired pneumonia: A prospective multi-centre study	3	Prediction model performance comparison study
75. (Blumentals et al., 2012) Body mass index and the incidence of influenza-associated pneumonia in a UK primary care cohort	4	Endpoints being influenza, influenza associated pneumonia.
76. (Afonso et al., 2012) The use of classification and regression trees to predict the likelihood of seasonal influenza	4	Endpoint being influenza
77. (Schepp et al., 2012) A clinical prediction rule for pneumonia after acute stroke	2	Endpoint being post stroke pneumonia in hospital settings
78. (Anonymous, 2010) Antipsychotic drugs in elderly patients associated with increased risk of pneumonia	8	No full text
79. (Thornton Snider et al., 2012) Inhaled corticosteroids and the risk of pneumonia in Medicare patients with COPD	4	Composite endpoint being outpatient and inpatient pneumonia

80. (Khan et al., 2011a) The association between statin use and the outcome of community acquired pneumonia: A systematic review and meta-analysis	4	Endpoint being CAP mortality
81. (Khan et al., 2011b) The association between statin use and the risk of community acquired pneumonia: A systematic review and meta-analysis	5	Abstract
82. (Wiersinga et al., 2012) SWAB/NVALT (Dutch working party on antibiotic policy and Dutch association of chest physicians) guidelines on the management of community-acquired pneumonia in adults	5	Guideline
83. (Woodhead et al., 2011b) Guidelines for the management of adult lower respiratory tract infections - Full version	5	Guideline
84. (Held et al., 2012) Validating and updating a risk model for pneumonia – a case study	3	Model validation and updating study
85. (Long et al., 2010) What factors predict pneumonia in patients with rheumatoid arthritis?	8	No full text
86. (Berg and Lindhardt, 2012) The role of procalcitonin in adult patients with community-acquired pneumonia - A systematic review	3	PCT diagnostic performance study
87. (Yende et al., 2010) The influence of pre-existing diabetes mellitus on the host immune response and outcome of pneumonia: analysis of two multicentre cohort studies	4	Endpoint being CAP mortality risk
88. (Watkins and Lemonovich, 2011) Diagnosis and management of community-acquired pneumonia in adults	5	Guidelines
89. (Viasus et al., 2011) Community-acquired pneumonia in patients with liver cirrhosis: Clinical features, outcomes, and usefulness of severity scores	2	Hospitalized CAP patients
90. (Vilanova et al., 2012) Obesity and metabolic syndrome as risk factors for community-acquired pneumonia	2	Emergency department setting
91. (Fung and Monteagudo-Chu, 2010) Community-acquired pneumonia in the elderly	5	Review

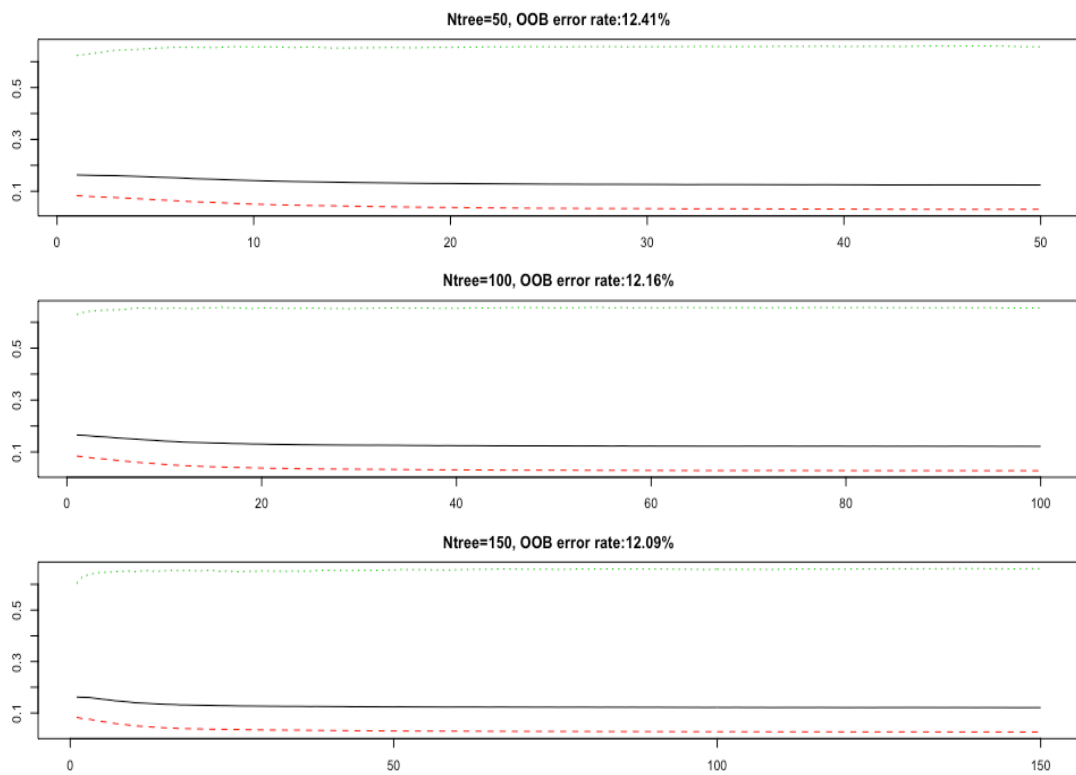
92. (Liapikou et al., 2012) Clinical presentation and evolution of community acquired pneumonia in older patients	8	No full text
93. (Hess et al., 2010) Comparative antibiotic failure rates in the treatment of community-acquired pneumonia: Results from a claims analysis	4	Endpoint being treatment failure
94. (Salluh et al., 2010) Cortisol levels and adrenal response in severe community-acquired pneumonia: a systematic review of the literature	4	Endpoint being severe CAP
95. (Lera et al., 2011) Differential features of Community Acquired Pneumonia (CAP) in young adults	2	Emergency department setting
96. (Viasus et al., 2010) Epidemiology, clinical features, and outcomes of community- acquired pneumonia in patients with liver cirrhosis	2	Hospitalized CAP
97. (Horie et al., 2012) Diagnostic and prognostic value of procalcitonin in community-acquired pneumonia	4	Endpoint being microbial aetiology and CAP outcome
98. (Schuetz et al., 2011) Prognostic value of procalcitonin in community-acquired pneumonia	4	Endpoint being CAP mortality risk
99. (Spindler et al., 2012) Swedish guidelines on the management of community-acquired pneumonia in immunocompetent adults--Swedish Society of Infectious Diseases 2012	5	Guidelines
100. (Sanz Herrero and Blanquer Olivas, 2012) Community-acquired pneumonia in adults	5	Seminar article
101. (Lin et al., 2019) Association of Increased Risk of Pneumonia and Using Proton Pump Inhibitors in Patients with Type II Diabetes Mellitus	4	Endpoint being hospitalized pneumonia
102. (Caldeira et al., 2012) Risk of pneumonia associated with use of angiotensin converting enzyme inhibitors and angiotensin receptor blockers: systematic review and meta-analysis	4	Endpoint being all cause pneumonia
103. (Dublin et al., 2011) Use of angiotensin-converting enzyme inhibitors is not associated with decreased pneumonia risk	8	No full text

104.	(He et al., 2013) Pneumonia and mortality risk in continuous ambulatory peritoneal dialysis patients with diabetic nephropathy	2	Peritoneal dialysis (PD) centre
105.	(Mapel et al., 2010) Pneumonia among COPD patients using inhaled corticosteroids and long-acting bronchodilators	4	Endpoint being all cause pneumonia
106.	(Nakanishi et al., 2010) Significance of the progression of respiratory symptoms for predicting community-acquired pneumonia in general practice	7	Patient age (groups) were not specified
107.	(Eom et al., 2011) Use of acid-suppressive drugs and risk of pneumonia: A systematic review and meta-analysis	4	Endpoint being CAP and HAP
108.	(Bonten et al., 2015) Polysaccharide conjugate vaccine against pneumococcal pneumonia in adults	4	Endpoint Being VE-CAP
109.	(Suaya et al., 2018) Post hoc analysis of the efficacy of the 13-valent pneumococcal conjugate vaccine against vaccine-type community-acquired pneumonia in at-risk older adults	4	Endpoint Being VE-CAP
110.	(Ochoa-Gondar et al., 2014) Effectiveness of the 23-valent pneumococcal polysaccharide vaccine against community-acquired pneumonia in the general population aged ≥ 60 years: 3 years of follow-up in the CAPAMIS study	4	Endpoint Being VE-CAP

Appendix F: Interim results during model development

Table A 10: General health status of study cohort frailty vs comorbidity

	No comorbidity	One comorbidity	Multi-comorbidity
Fit	42,473 (76.79%)	14,998 (47.42%)	2,774 (12.67%)
Mild	10,250 (18.53%)	11,159 (35.28%)	7,004 (31.98%)
Moderate	2,127 (3.85%)	4,194 (13.26%)	6,953 (31.75%)
Severe	463 (0.84%)	1,278 (4.04%)	5,169 (23.60%)
Total	55,313 (100%)	31,629 (100%)	21,900 (100%)



Green line: misclassification rate of pneumonia category

Black line: overall misclassification rate of both categories

Red line: misclassification rate of non-pneumonia category

Figure A 1: Number of trees for random forest 50, 100 and 150 for full model

Table A 11: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for full model

	Mtry5	Mtry10	Mtry15	Mtry20	Mtry25	Mtry30	Overall
Age	19	20	20	20	20	20	119
Chest Infection	20	19	19	19	19	19	115
eFrailty Index	18	18	18	18	18	18	108
BMI category	17	17	17	17	17	17	102
Season	15	16	16	16	16	16	95
Smoking Status	12	13	14	15	15	15	84
Charlson Score	14	14	13	14	14	14	83
Age group	16	15	15	13	12	10	81
Charlson Count	13	12	12	11	11	12	71
Antibiotic Ever	10	10	11	12	13	13	69
eFrailty category	11	11	9	9	9	9	58
Gender	3	9	10	10	10	11	53
Antibiotic	8	8	8	8	8	7	47
Immune system condition	6	7	7	7	7	8	42
Multi-comorbidity	7	6	6	5	3	3	30
Asthma drug	4	4	4	6	6	6	30
Cough	9	2	1	4	4	5	25
Flu vaccination	5	1	2	3	5	4	20
Pneumococcal Vaccination	2	5	5	2	1	1	16
Clinical test	1	3	3	1	2	2	12

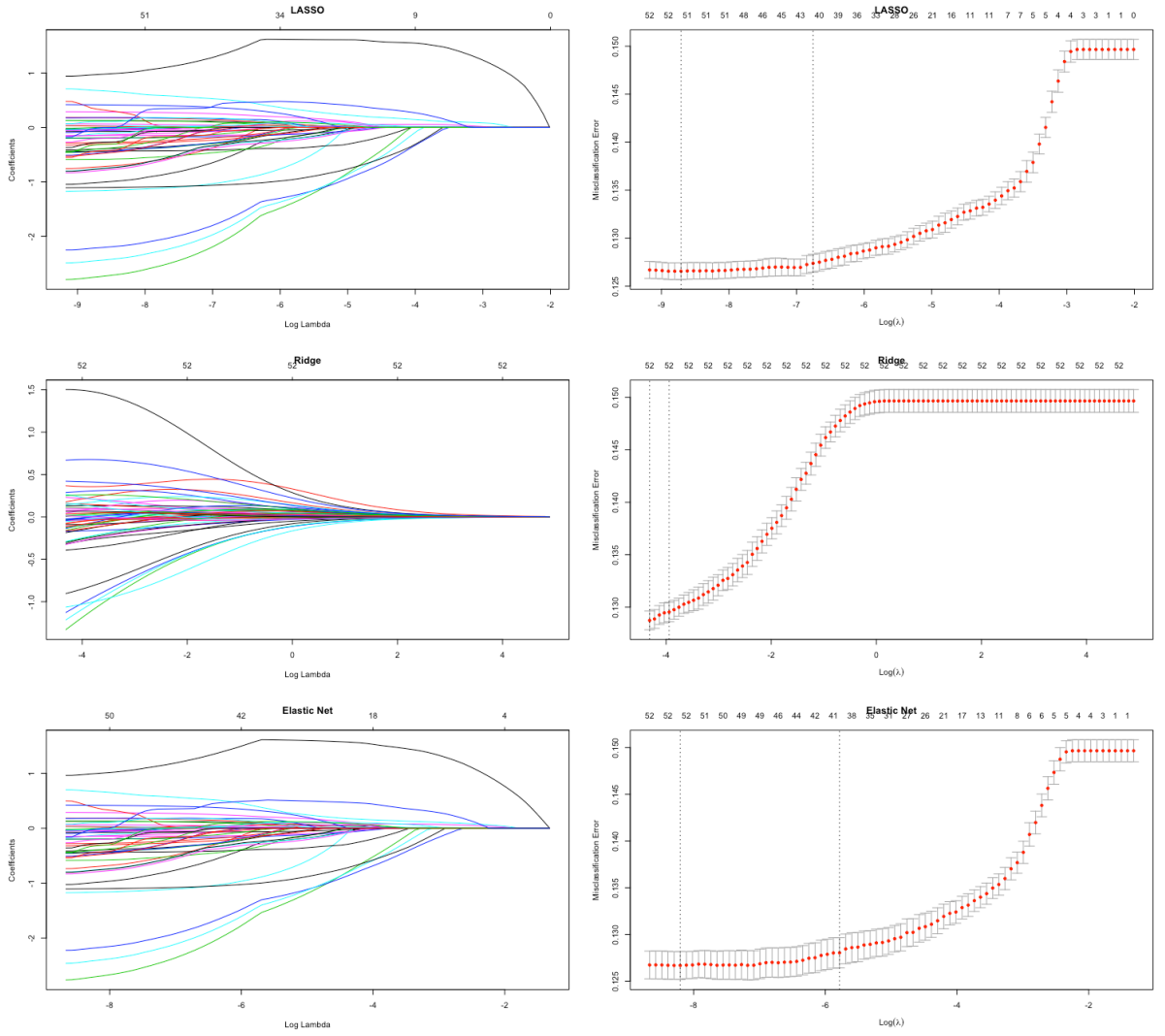


Figure A 2: lambda for 3 penalize regressions (full model)

Table A 12: Lambda and alpha for elastic net (full model)

alpha	mim lamda	mim lamda+1se	misclassification error (mim lamda)	misclassification error (mim lamda+1se)
0	0.014011	0.020637	0.128902	0.12961
0.01	0.002623	0.008458	0.127221	0.128232
0.02	0.002432	0.009431	0.127194	0.128259
0.03	0.002366	0.005326	0.127138	0.127772
0.04	0.002297	0.005567	0.127074	0.127837
0.05	0.00221	0.005873	0.126918	0.127947
0.06	0.002567	0.005724	0.127166	0.128094
0.07	0.00214	0.005481	0.127138	0.128149
0.08	0.002092	0.005766	0.127093	0.128112
0.09	0.002058	0.005126	0.127083	0.127919
0.1	0.001975	0.005599	0.12712	0.128066
0.11	0.001933	0.003686	0.126799	0.127644
0.12	0.00189	0.004538	0.12701	0.127855
0.13	0.002004	0.004387	0.127074	0.127901
0.14	0.002156	0.003388	0.126955	0.127506
0.15	0.001736	0.004286	0.126835	0.127837
0.16	0.001852	0.003404	0.127083	0.127616
0.17	0.001696	0.006222	0.127138	0.128461
0.18	0.001662	0.005117	0.126973	0.128278
0.19	0.001964	0.004107	0.12701	0.127901
0.2	0.001625	0.003493	0.126854	0.127607
0.21	0.001562	0.003648	0.126863	0.127616
0.22	0.001547	0.003482	0.126991	0.127726
0.23	0.001507	0.004391	0.126991	0.128195
0.24	0.001628	0.003251	0.126982	0.127772
0.25	0.00152	0.005279	0.127028	0.128553
0.26	0.001476	0.003142	0.12701	0.127681
0.27	0.001573	0.003111	0.127037	0.12769
0.28	0.001502	0.003607	0.127028	0.127919
0.29	0.001385	0.003326	0.12701	0.127929
0.3	0.001686	0.003305	0.126936	0.127818
0.31	0.001647	0.003573	0.1269	0.12802
0.32	0.001799	0.002878	0.126955	0.127763
0.33	0.001424	0.002923	0.126854	0.127681

0.34	0.001544	0.004217	0.126872	0.128342
0.35	0.00133	0.003253	0.127001	0.127938
0.36	0.001405	0.002884	0.12701	0.127717
0.37	0.001458	0.003135	0.126982	0.127883
0.38	0.001319	0.004884	0.12723	0.128572
0.39	0.001333	0.003057	0.127001	0.128057
0.4	0.001479	0.003519	0.126973	0.128241
0.41	0.001304	0.003562	0.127001	0.128076
0.42	0.001396	0.003085	0.127019	0.128002
0.43	0.001389	0.003041	0.127037	0.127929
0.44	0.001238	0.003083	0.127138	0.128103
0.45	0.001429	0.002724	0.126991	0.128039
0.46	0.001411	0.002568	0.12701	0.127809
0.47	0.001191	0.002378	0.127074	0.127699
0.48	0.001352	0.003337	0.126946	0.128213
0.49	0.001208	0.003209	0.127037	0.128223
0.5	0.001286	0.00232	0.126946	0.127607
0.51	0.001088	0.002837	0.12701	0.128131
0.52	0.001087	0.003377	0.127102	0.12825
0.53	0.001047	0.002292	0.12689	0.127598
0.54	0.001224	0.002607	0.126872	0.127818
0.55	0.001009	0.002356	0.126918	0.127745
0.56	0.001009	0.0024	0.126863	0.127671
0.57	0.001067	0.002231	0.126909	0.127726
0.58	0.000974	0.002339	0.126964	0.127965
0.59	0.00094	0.002321	0.12689	0.127984
0.6	0.000925	0.002526	0.126771	0.127846
0.61	0.000918	0.001955	0.126909	0.127579
0.62	0.00087	0.00209	0.126946	0.127708
0.63	0.000889	0.002095	0.127148	0.127625
0.64	0.000883	0.002457	0.127148	0.12802
0.65	0.001328	0.002126	0.126982	0.127818
0.66	0.000856	0.002449	0.127056	0.128158
0.67	0.000836	0.002325	0.126955	0.127919
0.68	0.001147	0.001786	0.126918	0.127469
0.69	0.000874	0.00204	0.126964	0.127892
0.7	0.00083	0.002125	0.126982	0.127873
0.71	0.001079	0.002451	0.127074	0.128269
0.72	0.00134	0.001955	0.127047	0.127809

0.73	0.000826	0.002001	0.127093	0.127855
0.74	0.001054	0.001956	0.127028	0.127837
0.75	0.001021	0.002385	0.127166	0.128149
0.76	0.001017	0.002166	0.126909	0.127846
0.77	0.000995	0.002571	0.127047	0.128232
0.78	0.00126	0.002336	0.127212	0.128204
0.79	0.001244	0.0026	0.127203	0.128259
0.8	0.001031	0.001727	0.127065	0.127754
0.81	0.000963	0.001923	0.126991	0.127947
0.82	0.001103	0.002264	0.127138	0.128085
0.83	0.0011	0.002385	0.127111	0.12837
0.84	0.001009	0.003051	0.12724	0.128636
0.85	0.000961	0.001937	0.127065	0.127956
0.86	0.000959	0.001845	0.127129	0.127809
0.87	0.000993	0.00191	0.126991	0.12803
0.88	0.000981	0.001837	0.127028	0.127938
0.89	0.000927	0.00265	0.127138	0.128553
0.9	0.000933	0.002241	0.127184	0.128324
0.91	0.000923	0.002078	0.127111	0.128103
0.92	0.000896	0.002233	0.126927	0.128232
0.93	0.000903	0.001787	0.127047	0.127919
0.94	0.000919	0.002247	0.127212	0.12836
0.95	0.000892	0.001717	0.127102	0.127855
0.96	0.000891	0.002062	0.127047	0.128158
0.97	0.000985	0.001811	0.127184	0.128085
0.98	0.000993	0.001912	0.127093	0.128103
0.99	0.001002	0.001742	0.127074	0.128057
1	0.000921	0.001631	0.127083	0.127873

Blue highlighted lambda and alpha are chosen for the elastic model

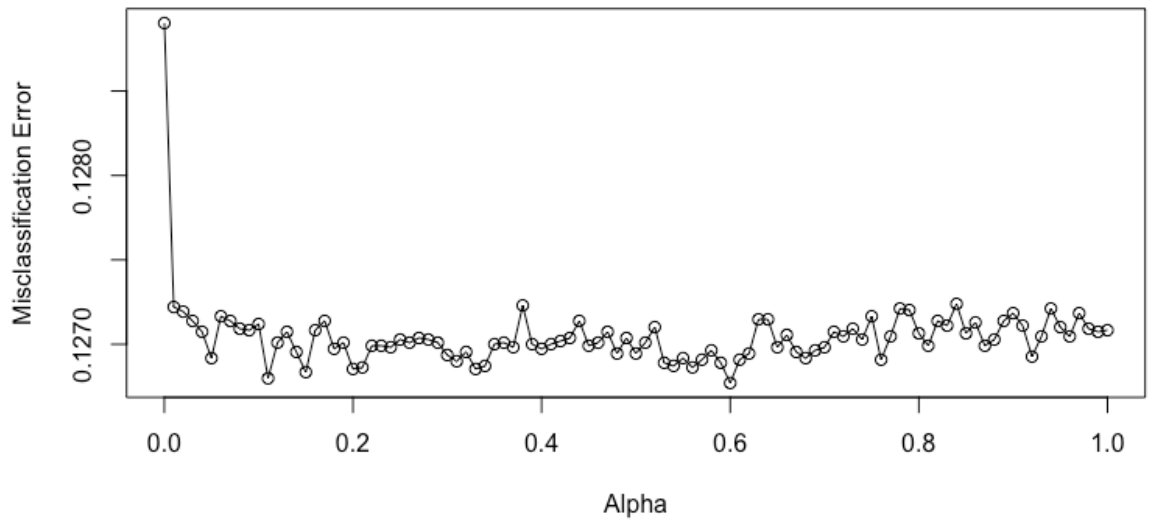


Figure A 3: Alpha for elastic net model against misclassification error (full model)

Table A 13: Absolute value of coefficients of three penalized regression models (full model)

	Lasso model		Ridge model		Elastic net model	
Variable	Importance	Sign	Importance	Sign	Importance	Sign
Age	0.03	POS	0.04	POS	0.02	POS
Age group (35,45]	0.11	POS	0.01	NEG	0.17	POS
Age group (45,55]	0.20	NEG	0.41	NEG	0.05	NEG
Age group (55,65]	0.27	NEG	0.56	NEG	0.06	NEG
Age group (65,75]	0.42	NEG	0.80	NEG	0.16	NEG
Age group (75,85]	0.30	NEG	0.75	NEG	0.00	NEG
Age group (85,110]	0.00	NEG	0.52	NEG	0.35	POS
Antibiotic (Yes)	0.15	NEG	0.15	NEG	0.14	NEG
Antibiotic prescription in the following 30 days after initial RTI consultations	0.07	NEG	0.08	NEG	0.07	NEG
Asthma drug (Yes)	1.10	NEG	1.11	NEG	1.07	NEG
Healthy weight	0.44	NEG	0.56	NEG	0.19	NEG
Overweight	0.73	NEG	0.85	NEG	0.48	NEG
Obese	0.76	NEG	0.88	NEG	0.51	NEG
Severe obese	0.73	NEG	0.86	NEG	0.47	NEG
Morbid obese	0.45	NEG	0.58	NEG	0.20	NEG
BMI information not recorded	0.37	NEG	0.48	NEG	0.12	NEG
Cancer (Yes)	0.05	NEG	0.05	NEG	0.01	NEG
Charlson Count	0.66	POS	0.74	POS	0.55	POS
Charlson Score	0.03	NEG	0.07	NEG	0.00	NEG
Chest Infection (Yes)	1.00	POS	0.93	POS	1.21	POS
Chronic heart disease (Yes)	0.30	NEG	0.34	NEG	0.25	NEG
Chronic kidney disease (Yes)	0.19	NEG	0.21	NEG	0.16	NEG
Chronic liver disease (Yes)	0.18	POS	0.17	POS	0.18	POS
Chronic neurological condition (Yes)	0.02	NEG	0.07	NEG	0.01	POS

Chronic respiratory disease (yes)	0.13	NEG	0.18	NEG	0.09	NEG
Clinical check (Yes)	1.15	NEG	1.19	NEG	1.09	NEG
Clinical test (Yes)	0.28	POS	0.29	POS	0.27	POS
Cold/ Influenza/ URTI (Yes)	0.98	NEG	1.06	NEG	0.74	NEG
Cough (Yes)	0.70	NEG	0.77	NEG	0.46	NEG
Diabetes (Yes)	0.39	NEG	0.44	NEG	0.32	NEG
eFrailty index	0.29	POS	0.76	POS	0.00	POS
Frailty (Mild)	0.00	NEG	0.05	NEG	0.01	POS
Frailty (Moderate)	0.05	NEG	0.16	NEG	0.00	NEG
Frailty (Severe)	0.10	NEG	0.27	NEG	0.00	NEG
Flu vaccination (Yes)	0.06	POS	0.07	POS	0.05	POS
Female	0.19	NEG	0.20	NEG	0.18	NEG
Hemiplegia (Yes)	0.02	POS	0.04	POS	0.03	POS
Immune system condition (Yes)	0.44	NEG	0.45	NEG	0.43	NEG
Hospital admission in previous year (Yes)	0.41	POS	0.43	POS	0.39	POS
Multi-comorbidity (one comorbidity)	0.27	NEG	0.29	NEG	0.22	NEG
Multi-comorbidity (more than one comorbidity)	0.58	NEG	0.59	NEG	0.50	NEG
Otitis media (Yes)	2.70	NEG	2.82	NEG	2.32	NEG
Peptic ulcer (Yes)	0.43	NEG	0.48	NEG	0.34	NEG
Pneumococcal vaccination (Yes)	0.12	POS	0.13	POS	0.12	POS
PVD (Yes)	0.42	NEG	0.48	NEG	0.33	NEG
Rhinosinusitis (Yes)	2.41	NEG	2.51	NEG	2.08	NEG
Season (spring)	0.03	NEG	0.05	NEG	0.01	NEG
Season (summer)	0.02	NEG	0.04	NEG	0.00	NEG
Season (winter)	0.03	NEG	0.04	NEG	0.01	NEG
Past smoker	0.18	POS	0.19	POS	0.16	POS
Current smoker	0.12	POS	0.13	POS	0.11	POS
Sore throat (Yes)	2.18	NEG	2.26	NEG	1.89	NEG

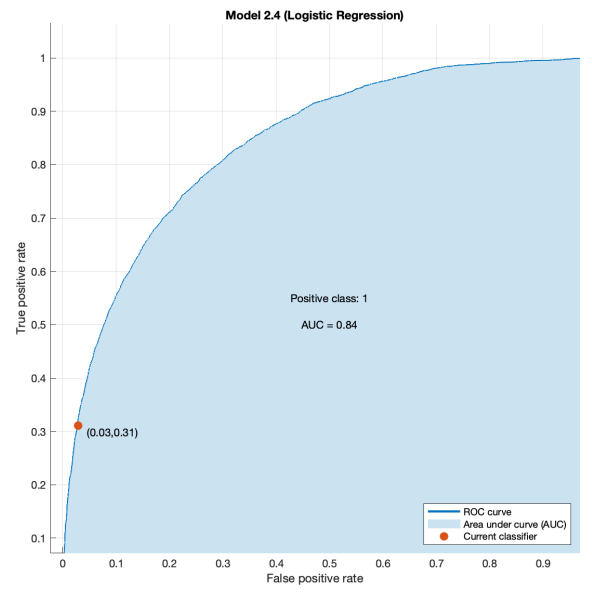
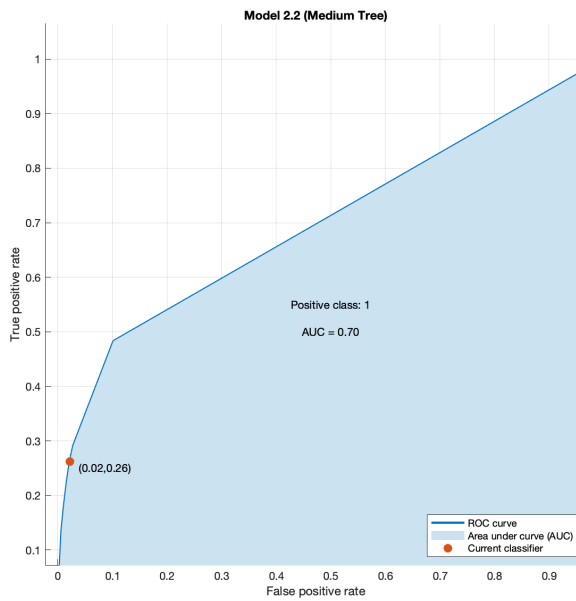
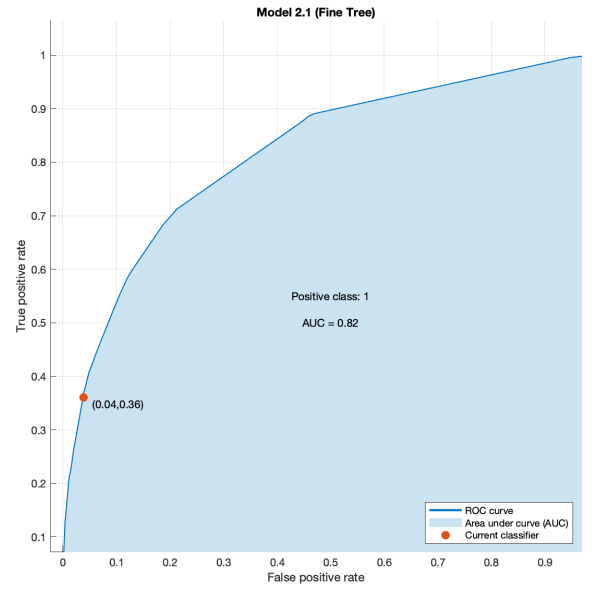
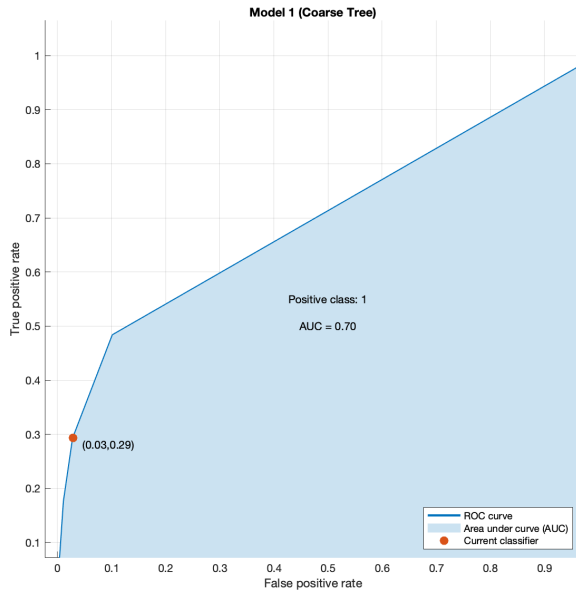
Table A 14: Variable importance ranking based on top 20 variables across three penalized regression models (full model)

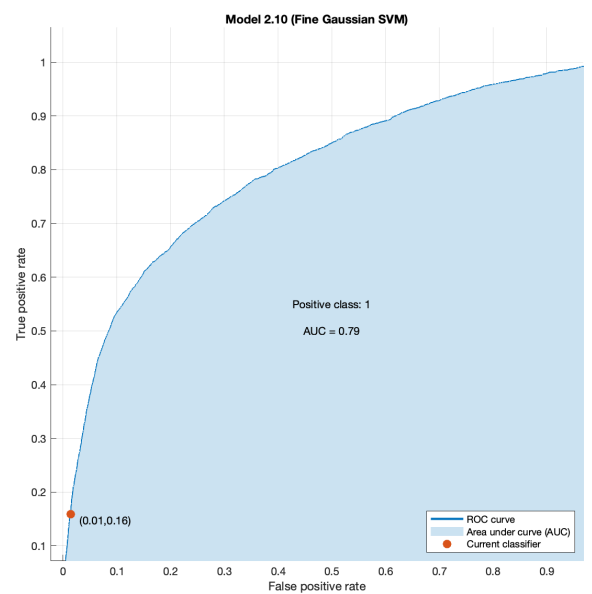
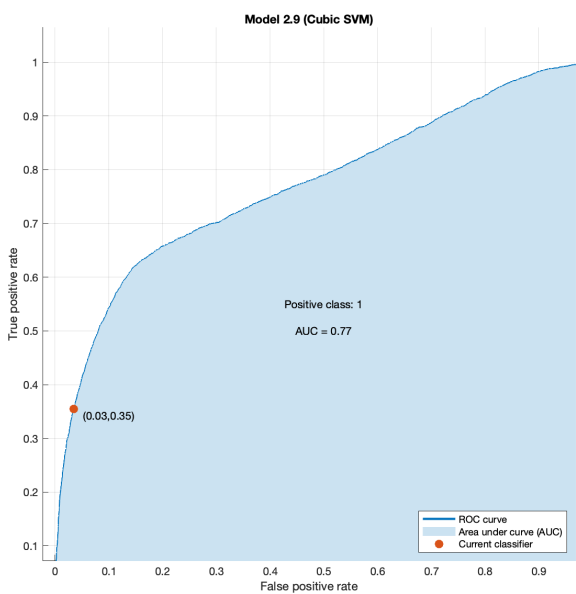
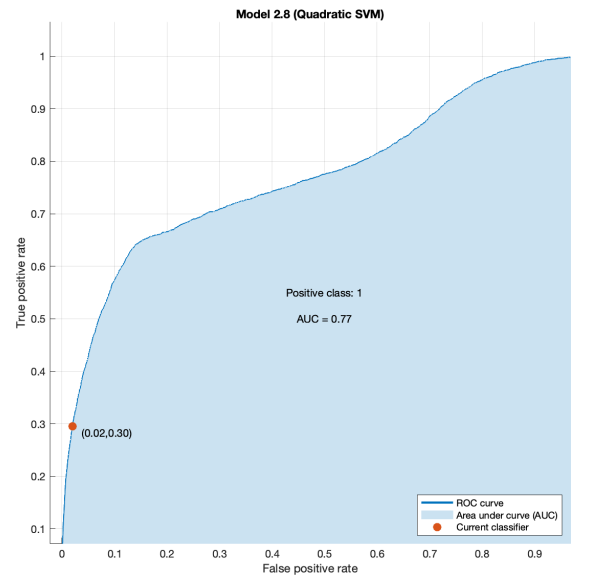
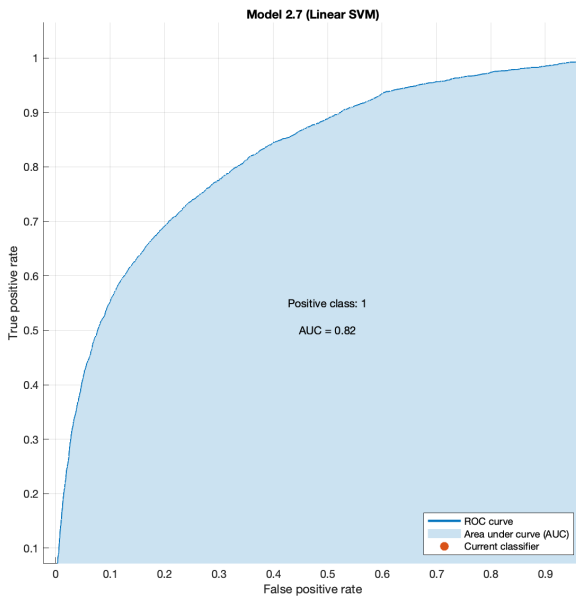
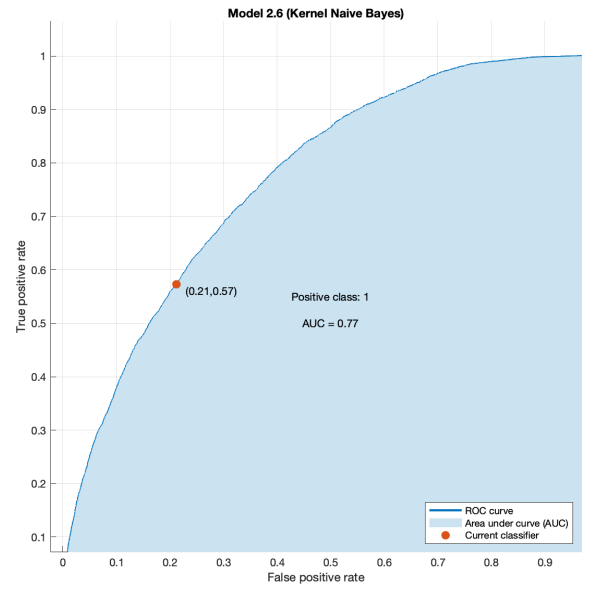
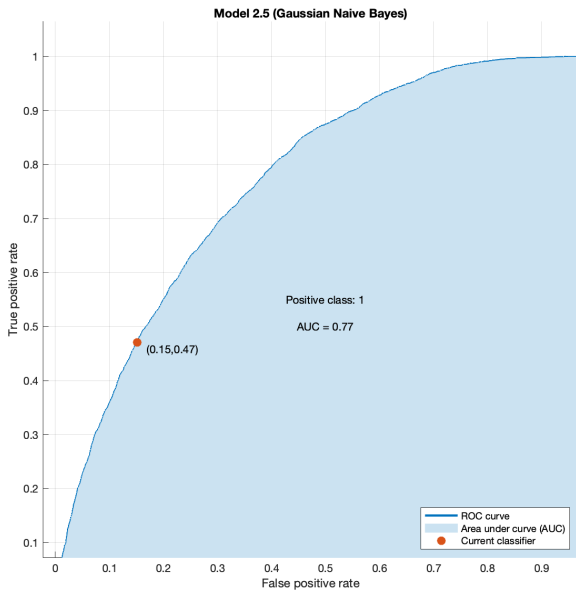
	Lasso model	Ridge model	Elastic net model	Overall
Chest infection	20	20	20	60
Charlson Count	19	18	19	56
Admission during previous year	18	17	18	53
Clinical test	16	16	16	48
Chronic liver disease	15	14	15	44
Smoking status	14	15	13	42
eFrailty index	17	19	5	41
Pneumococcal vaccination	13	13	12	38
Age group	11	8	17	36
Hemiplegia	8	10	9	27
Age	9	9	8	26
Flu vaccination	0	11	10	21
eFrailty category	7	4	6	17
Season	4	7	4	15
Chronic neurological disease	5	2	7	14
Flu vaccination	10	0	0	10
Charlson Score	1	1	1	3
cancer	0	3	0	3

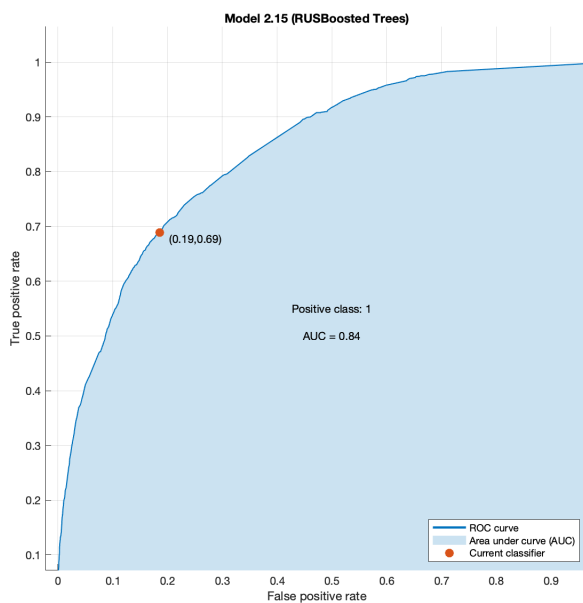
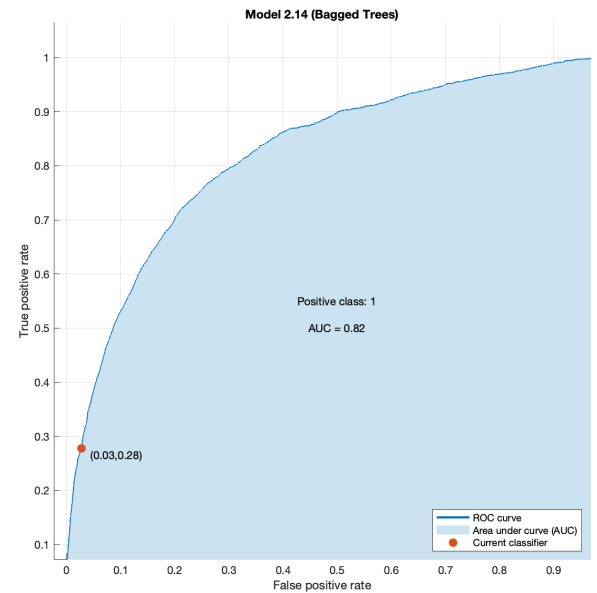
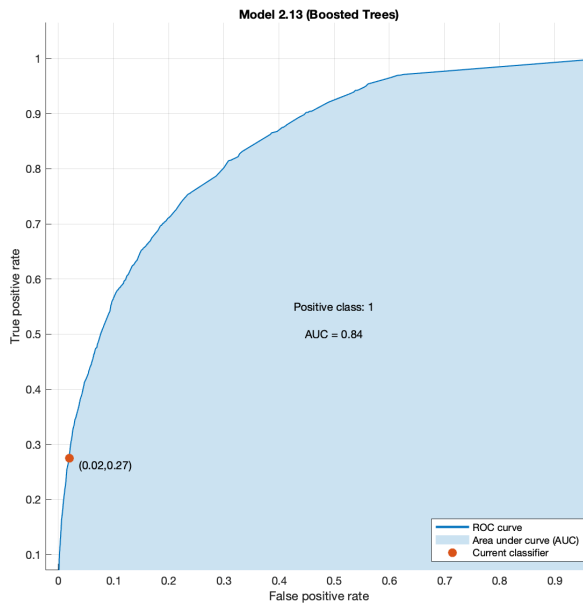
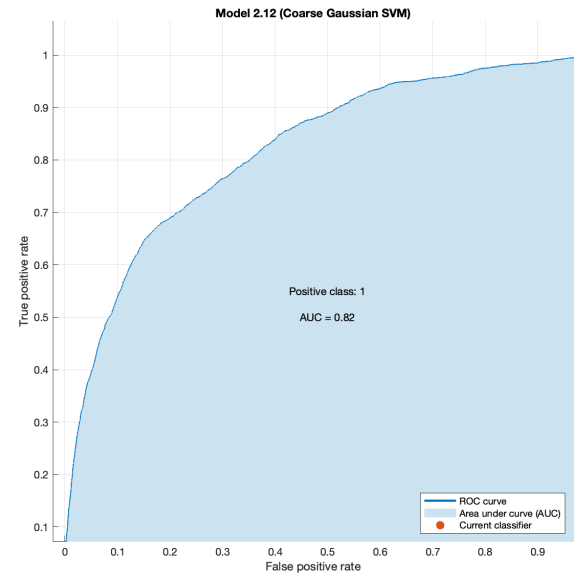
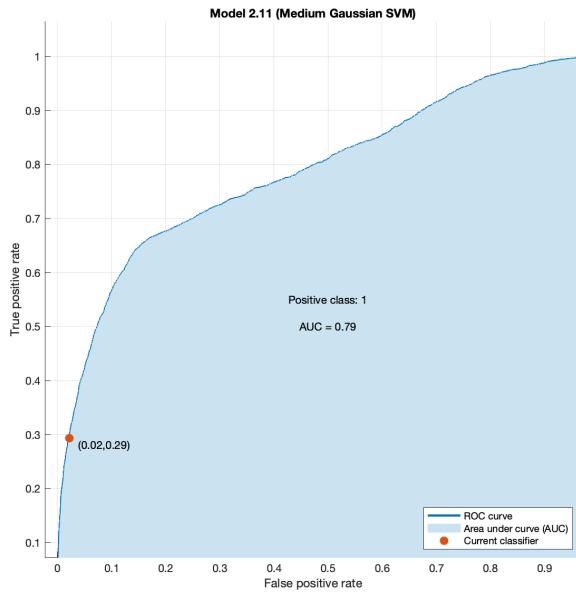
Table A 15: Machine learning model comparison using full dataset with 10-fold cross validation

Decision Trees	Accuracy	AUC	Sensitivity	Specificity
Coarse Tree (Max number of split: 4)	87.0%	0.70	29.3%	97.1%
Fine Trees	87.2%	0.82	36.1%	96.2%
Medium Tree (Max number of split: 20)	87.1%	0.70	26.3%	97.8%
Logistic Regression	87.3%	0.84	31.1%	97.2%
Naïve Bayes				
Gaussian Naïve Bayes	79.2%	0.77	47.1%	84.8%
Kernel Naïve Bayes	75.6%	0.77	57.3%	78.8%
Support Vector Machines (SVM)				
Linear SVM	85%	0.82	0%	100%
Quadratic SVM	87.7%	0.77	29.5%	98%
Cubic SVM	87.4%	0.77	35.4%	96.5%
Fine Gaussian SVM	86.3%	0.79	15.9%	98.6%
Medium Gaussian SVM	87.6%	0.79	29.3%	97.8%
Coarse Gaussian SVM	85.8%	0.82	6.7%	99.7%
Ensemble Classifiers				
Boosted Trees	87.4%	0.84	27.5%	96%
Bagged Trees	86.9%	0.82	27.8%	97.3%
RUSBoost Trees	79.5%	0.84	68.9%	81.4%

Table A 16: Machine learning model using full dataset with 10-fold cross validation ROC curves







Note: The red dot on the plots shows the performance of the currently selected classifier. This marker shows the values of the false positive rate (FPR) and the true positive rate (TPR) for the currently selected classifier. For example, a false positive rate (FPR) of 0.2 indicates that the current classifier assigns 20% of the observations incorrectly to the positive class. A true positive rate of 0.9 indicates that the current classifier assigns 90% of the observations correctly to the positive class.

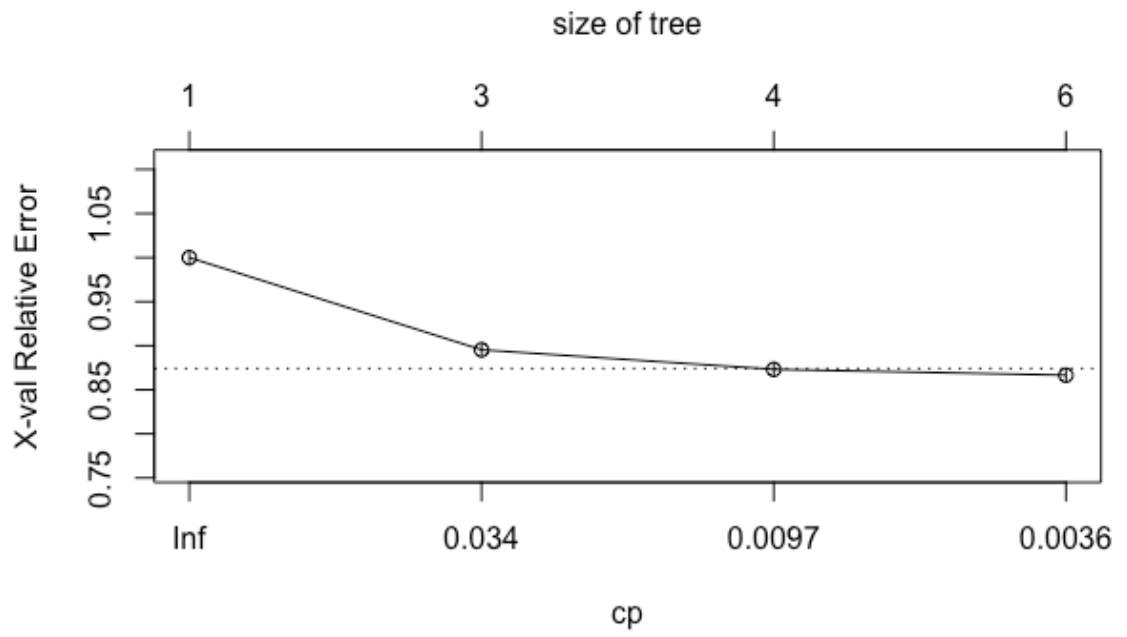
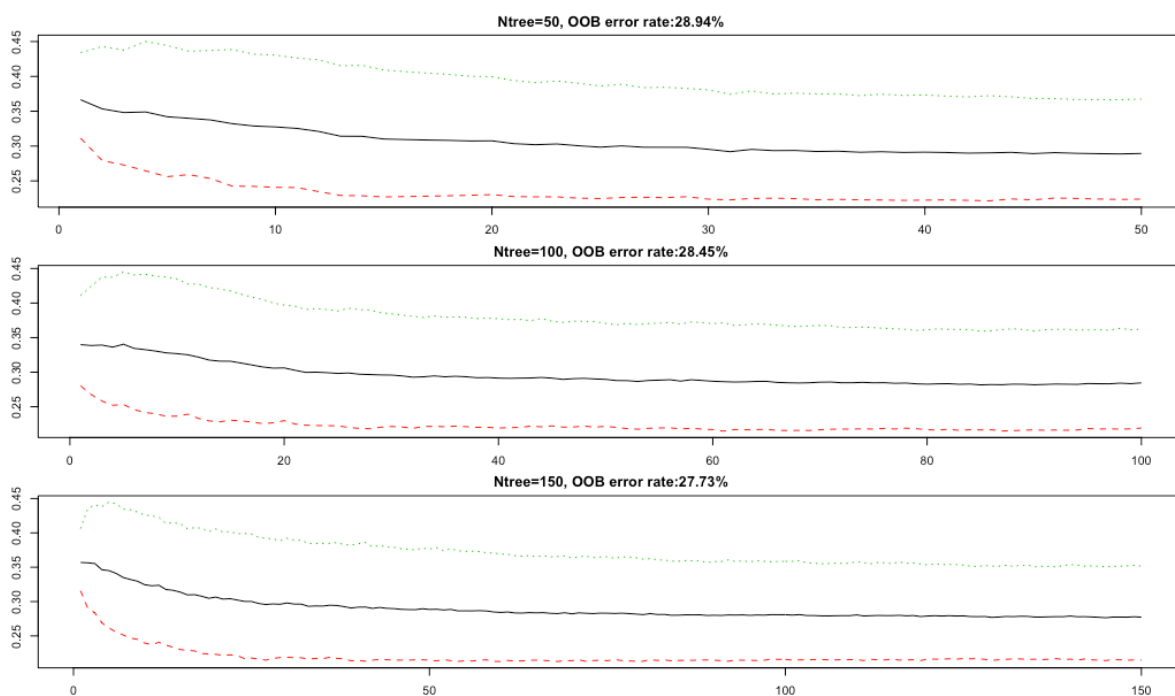


Figure A 4: Tuning parameter for CART full model

Table A 17: Comparison statistics of model variables for development data and temporal validation data (full model). Figures are frequencies (column percentages) except where indicated.

	Development data		Temporal validation data	
	Non-pneumonia patients (n=92,553)	Pneumonia patients (n=16,289)	Non-pneumonia patients (n=18,022)	Pneumonia patients (n=3,310)
Age group				
16-35	26,518 (28.7)	1,446 (8.9)	5059 (28.1)	263 (7.9)
36-45	15,640 (16.9)	1,598 (9.8)	2672 (14.8)	253 (7.6)
46-55	14,616 (15.8)	1,725 (10.6)	2935 (16.3)	365 (11.0)
56-65	13,880 (15.0)	2,480 (15.2)	2736 (15.2)	456 (13.8)
66-75	11,607 (12.5)	3,020 (18.5)	2536 (14.1)	687 (20.8)
76-85	7,669 (8.3)	3,549 (21.8)	1533 (8.5)	732 (22.1)
86 and above	2,623 (2.8)	2,471 (15.2)	551 (3.1)	554 (16.7)
Charlson Comorbidity Count				
0	52,182 (56.4)	5,142 (31.6)	9914 (55.0)	941 (28.4)
1	28,523 (30.8)	5,027 (30.9)	5574 (30.9)	1006 (30.4)
2	7,434 (8.0)	2,964 (18.2)	1544 (8.6)	624 (18.9)
3	2,866 (3.1)	1,734 (10.6)	645 (3.6)	370 (11.2)
4	1,046 (1.1)	878 (5.4)	241 (1.3)	221 (6.7)
5	364 (0.4)	371 (2.3)	77 (0.4)	104 (3.1)
6	138 (0.1)	173 (1.1)	27 (0.1)	44 (1.3)
Antibiotic prescription on the RTI index date	49,591 (53.6)	9,114 (56.0)	8708 (48.3)	1863 (56.3)
Chest Infection	9,461 (10.2)	7,942 (48.8)	1,338 (7.4)	1,404 (42.4)



Green line: misclassification rate of pneumonia category

Black line: overall misclassification rate of both categories

Red line: misclassification rate of non-pneumonia category

Figure A 5: Number of trees for random forest 50, 100 and 150 for LRTI model

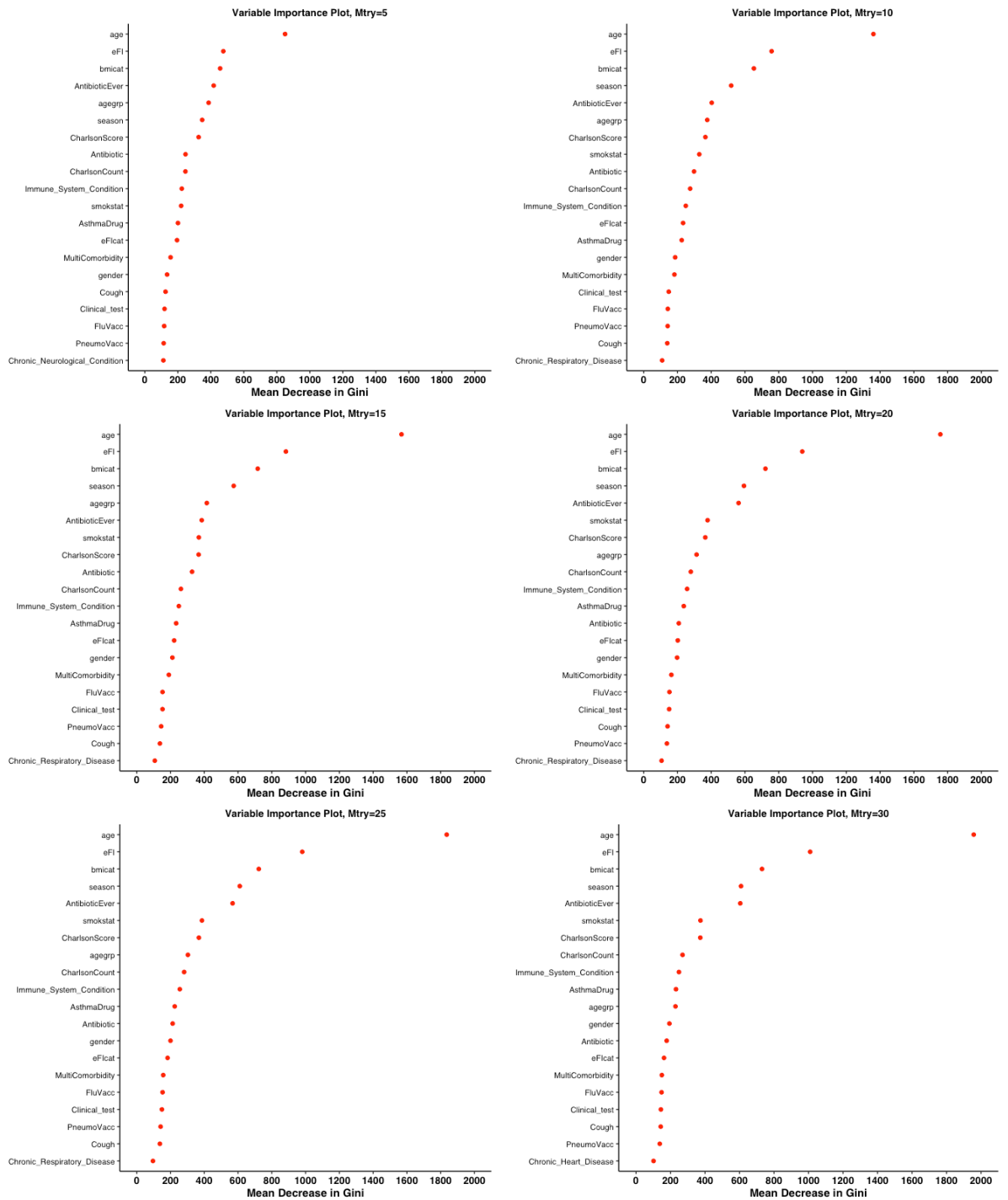


Figure A 6: Variable importance results by random forest models with Mtry 5-30 (5 increments) for LRTI model

Table A 18: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for LRTI model

	Mtry5	Mtry10	Mtry15	Mtry20	Mtry25	Mtry30	Overall
Age	20	20	20	20	20	20	120
eFrailty Index	19	19	19	19	19	19	114
BMI category	18	18	18	18	18	18	108
Season	15	17	17	17	17	17	100
Antibiotic Ever	17	16	15	16	16	16	96
Age group	16	15	16	13	13	10	83
Charlson Score	14	14	13	14	14	14	83
Smoking Status	10	13	14	15	15	15	82
Charlson Count	12	11	11	12	12	13	71
Immune system condition	11	10	10	11	11	12	65
Antibiotic	13	12	12	9	9	8	63
Asthma drug	9	8	9	10	10	11	57
eFrailty category	8	9	8	8	7	7	47
Gender	6	7	7	7	8	9	44
Multi-comorbidity	7	6	6	6	6	6	37
Flu vaccination	3	4	5	5	5	5	27
Clinical test	4	5	4	4	4	4	25
Cough	5	2	2	3	2	3	17
Pneumococcal Vaccination	2	3	3	2	3	2	15
Chronic respiratory disease	0	1	1	1	1	0	4

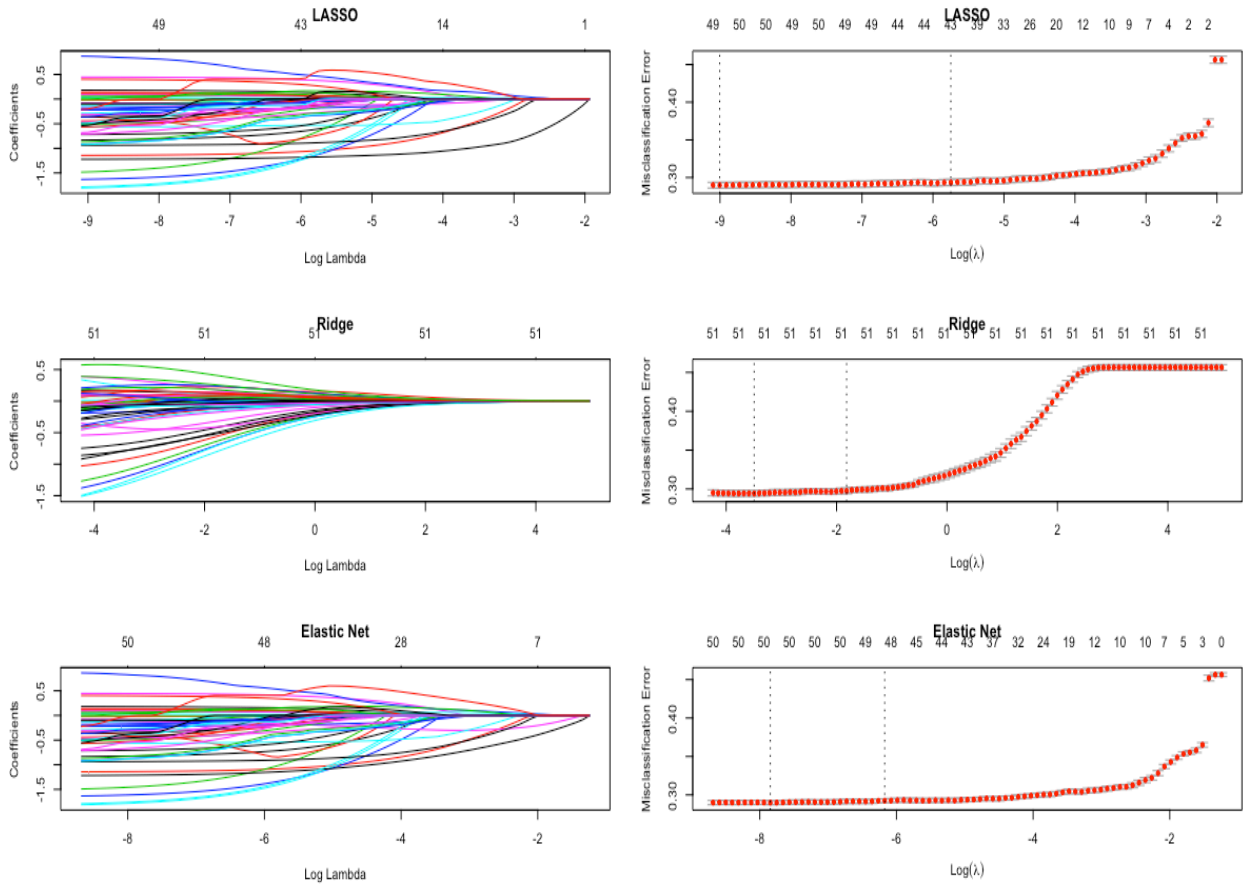


Figure A 7: Lambda for 3 penalize regressions (LRTI model)

Table A 19: Lambda and alpha for elastic net (LRTI model)

alpha	mim lamda	mim lamda+1se	misclassification error (mim lamda)	misclassification error (mim lamda+1se)
0	0.014452	0.105877	0.293455	0.296156
0.01	0.007057	0.054641	0.292938	0.295351
0.02	0.006366	0.018043	0.291789	0.293915
0.03	0.005545	0.069981	0.291559	0.296213
0.04	0.003935	0.049205	0.290984	0.294719
0.05	0.004427	0.030974	0.290812	0.293455
0.06	0.003828	0.037669	0.291272	0.294777
0.07	0.004019	0.028906	0.291157	0.294087
0.08	0.004113	0.023931	0.290984	0.293857
0.09	0.004397	0.038377	0.291272	0.294604
0.1	0.003921	0.041918	0.290927	0.295984
0.11	0.004247	0.035073	0.290755	0.294892
0.12	0.004269	0.031855	0.29041	0.294547
0.13	0.004567	0.032543	0.291616	0.294949
0.14	0.004012	0.010088	0.290697	0.293225
0.15	0.004587	0.026933	0.291387	0.294834
0.16	0.003815	0.025018	0.291387	0.294834
0.17	0.003492	0.020506	0.290755	0.29403
0.18	0.004115	0.012327	0.291502	0.29449
0.19	0.003555	0.026775	0.291329	0.295696
0.2	0.003285	0.022151	0.291789	0.294892
0.21	0.003217	0.017383	0.290582	0.294834
0.22	0.003245	0.011797	0.290352	0.29334
0.23	0.00328	0.022739	0.292191	0.295984
0.24	0.004463	0.019509	0.291042	0.295122
0.25	0.003074	0.010286	0.292019	0.294604
0.26	0.003153	0.021456	0.291846	0.296845
0.27	0.003454	0.01955	0.291789	0.295811
0.28	0.002695	0.010353	0.290984	0.294202
0.29	0.0051	0.009546	0.291099	0.293743
0.3	0.003378	0.009574	0.290984	0.294375
0.31	0.003239	0.006233	0.290812	0.293225
0.32	0.003347	0.012742	0.291731	0.294432
0.33	0.003368	0.010275	0.291272	0.293283

0.34	0.002668	0.008605	0.291444	0.293743
0.35	0.002545	0.014535	0.291904	0.295007
0.36	0.002841	0.014002	0.290755	0.294949
0.37	0.00243	0.008055	0.290869	0.293455
0.38	0.002432	0.013512	0.291559	0.294949
0.39	0.002797	0.012457	0.291157	0.295466
0.4	0.002289	0.012718	0.290755	0.295122
0.41	0.002233	0.008581	0.291444	0.29403
0.42	0.002262	0.007431	0.291272	0.294202
0.43	0.002189	0.006804	0.29064	0.293168
0.44	0.002062	0.007496	0.290755	0.29403
0.45	0.002054	0.004752	0.290697	0.29311
0.46	0.002028	0.00972	0.291789	0.294949
0.47	0.002665	0.007417	0.291272	0.294087
0.48	0.002035	0.010697	0.290812	0.295754
0.49	0.002068	0.012835	0.292019	0.295869
0.5	0.003214	0.009897	0.291214	0.29449
0.51	0.002043	0.010161	0.291616	0.295064
0.52	0.002004	0.005838	0.291329	0.29334
0.53	0.001826	0.008514	0.290122	0.293513
0.54	0.002278	0.006165	0.290927	0.292938
0.55	0.001843	0.012197	0.290525	0.295466
0.56	0.001826	0.005836	0.290927	0.292823
0.57	0.001745	0.006894	0.290352	0.293743
0.58	0.001715	0.008771	0.290582	0.29449
0.59	0.001799	0.005051	0.290295	0.292593
0.6	0.001785	0.006796	0.290984	0.294432
0.61	0.002997	0.012171	0.291904	0.297363
0.62	0.00168	0.006456	0.290352	0.293685
0.63	0.002102	0.006012	0.290352	0.293283
0.64	0.001922	0.007661	0.291444	0.29449
0.65	0.001648	0.008426	0.291559	0.295811
0.66	0.001699	0.007293	0.290352	0.294375
0.67	0.001584	0.010293	0.291674	0.296845
0.68	0.001504	0.010236	0.291731	0.295581
0.69	0.001469	0.005695	0.290525	0.29449
0.7	0.001475	0.005361	0.291042	0.293972
0.71	0.001882	0.007857	0.292076	0.294892
0.72	0.001447	0.005611	0.291157	0.293915

0.73	0.001551	0.007784	0.292019	0.296041
0.74	0.001434	0.004843	0.290467	0.293685
0.75	0.001552	0.005589	0.290869	0.293857
0.76	0.001396	0.006275	0.290697	0.29449
0.77	0.002206	0.007657	0.291272	0.294892
0.78	0.001697	0.006403	0.290927	0.2938
0.79	0.001394	0.004292	0.290984	0.293225
0.8	0.001441	0.008781	0.291502	0.296673
0.81	0.001561	0.006951	0.29064	0.29449
0.82	0.001343	0.005255	0.291789	0.293915
0.83	0.002415	0.005693	0.291846	0.29449
0.84	0.001411	0.006891	0.291616	0.294834
0.85	0.001677	0.005876	0.291559	0.295064
0.86	0.002011	0.006607	0.291444	0.294834
0.87	0.001301	0.008455	0.291502	0.296098
0.88	0.001398	0.006457	0.291444	0.295294
0.89	0.001226	0.00681	0.291214	0.295409
0.9	0.001281	0.00725	0.291444	0.296041
0.91	0.001244	0.004959	0.291329	0.29357
0.92	0.001325	0.00845	0.291674	0.296156
0.93	0.001275	0.005943	0.290525	0.293857
0.94	0.001333	0.007269	0.291789	0.295581
0.95	0.001331	0.005927	0.291616	0.295122
0.96	0.001526	0.005601	0.290122	0.293398
0.97	0.001352	0.00461	0.291559	0.294547
0.98	0.001625	0.0063	0.292191	0.294777
0.99	0.001388	0.003332	0.291042	0.293513
1	0.001349	0.005842	0.291502	0.294547

Blue highlighted lambda and alpha are chosen for the elastic model

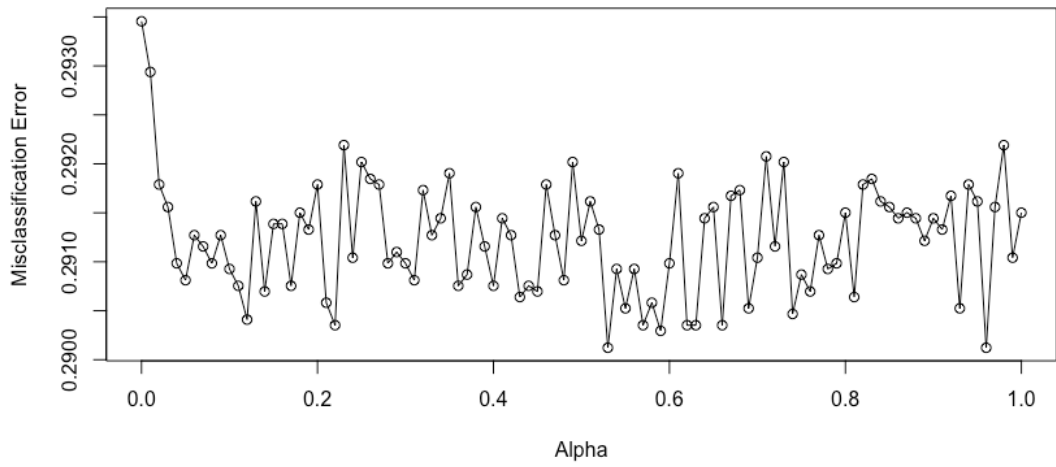


Figure A 8: Alpha for elastic net model against misclassification error (LRTI model)

Table A 20: Absolute value of coefficients of three penalized regression models (LRTI model)

	Lasso model		Ridge model		Elastic net model	
Variable	Importance	Sign	Importance	Sign	Importance	Sign
Age	0.03	POS	0.04	POS	0.02	POS
Age group (35,45]	0.04	POS	0.05	NEG	0.12	POS
Age group (45,55]	0.28	NEG	0.45	NEG	0.11	NEG
Age group (55,65]	0.38	NEG	0.62	NEG	0.14	NEG
Age group (65,75]	0.53	NEG	0.84	NEG	0.22	NEG
Age group (75,85]	0.36	NEG	0.74	NEG	0.00	NEG
Age group (85,110]	0.00	NEG	0.44	NEG	0.41	POS
Antibiotic (Yes)	0.12	NEG	0.12	NEG	0.15	NEG
Antibiotic prescription in the following 30 days after initial RTI consultations	1.22	NEG	1.22	NEG	1.17	NEG
Asthma drug (Yes)	0.93	NEG	0.94	NEG	0.92	NEG
Healthy weight	0.51	NEG	0.59	NEG	0.23	NEG
Overweight	0.82	NEG	0.90	NEG	0.52	NEG
Obese	0.86	NEG	0.94	NEG	0.57	NEG
Severe obese	0.87	NEG	0.95	NEG	0.57	NEG
Morbid obese	0.66	NEG	0.74	NEG	0.36	NEG
BMI information not recorded	0.44	NEG	0.52	NEG	0.15	NEG
Cancer (Yes)	0.09	NEG	0.09	NEG	0.11	NEG
Charlson Count	0.83	POS	0.88	POS	0.59	POS
Charlson Score	0.12	NEG	0.15	NEG	0.00	NEG
Chest Infection (Yes)	0.46	NEG	0.48	NEG	0.38	NEG
Chronic heart disease (Yes)	0.10	NEG	0.09	NEG	0.10	NEG
Chronic kidney disease (Yes)	0.11	POS	0.12	POS	0.10	POS
Chronic liver disease (Yes)	0.06	POS	0.04	POS	0.12	POS
Chronic neurological condition (Yes)	0.17	NEG	0.19	NEG	0.10	NEG

Chronic respiratory disease (yes)	1.74	NEG	1.81	NEG	1.51	NEG
Clinical check (Yes)	0.44	POS	0.45	POS	0.42	POS
Clinical test (Yes)	0.82	NEG	0.84	NEG	0.75	NEG
Cold/ Influenza/ URTI (Yes)	1.14	NEG	1.15	NEG	1.11	NEG
Cough (Yes)	0.35	NEG	0.37	NEG	0.25	NEG
Diabetes (Yes)	0.43	NEG	0.32	NEG	0.69	NEG
eFrailty index	0.09	NEG	0.11	NEG	0.03	NEG
Frailty (Mild)	0.19	NEG	0.22	NEG	0.09	NEG
Frailty (Moderate)	0.25	NEG	0.31	NEG	0.10	NEG
Frailty (Severe)	0.09	POS	0.09	POS	0.07	POS
Flu vaccination (Yes)	0.22	NEG	0.22	NEG	0.20	NEG
Female	0.00	NEG	0.00	POS	0.00	NEG
Hemiplegia (Yes)	0.56	NEG	0.57	NEG	0.55	NEG
Immune system condition (Yes)	0.39	POS	0.41	POS	0.36	POS
Hospital admission in previous year (Yes)	0.30	NEG	0.32	NEG	0.25	NEG
Multi-comorbidity (one comorbidity)	0.71	NEG	0.72	NEG	0.62	NEG
Multi-comorbidity (more than one comorbidity)	1.43	NEG	1.52	NEG	1.17	NEG
Otitis media (Yes)	0.42	NEG	0.45	NEG	0.28	NEG
Peptic ulcer (Yes)	0.18	POS	0.18	POS	0.16	POS
Pneumococcal vaccination (Yes)	0.49	NEG	0.52	NEG	0.35	NEG
PVD (Yes)	1.76	NEG	1.84	NEG	1.55	NEG
Rhinosinusitis (Yes)	0.02	NEG	0.03	NEG	0.00	NEG
Season (spring)	0.08	POS	0.08	POS	0.08	POS
Season (summer)	0.10	NEG	0.10	NEG	0.08	NEG
Season (winter)	0.13	POS	0.14	POS	0.12	POS
Past smoker	0.00	NEG	0.00	NEG	0.00	NEG
Current smoker	1.60	NEG	1.65	NEG	1.44	NEG
Sore throat (Yes)	2.18	NEG	2.26	NEG	1.89	NEG

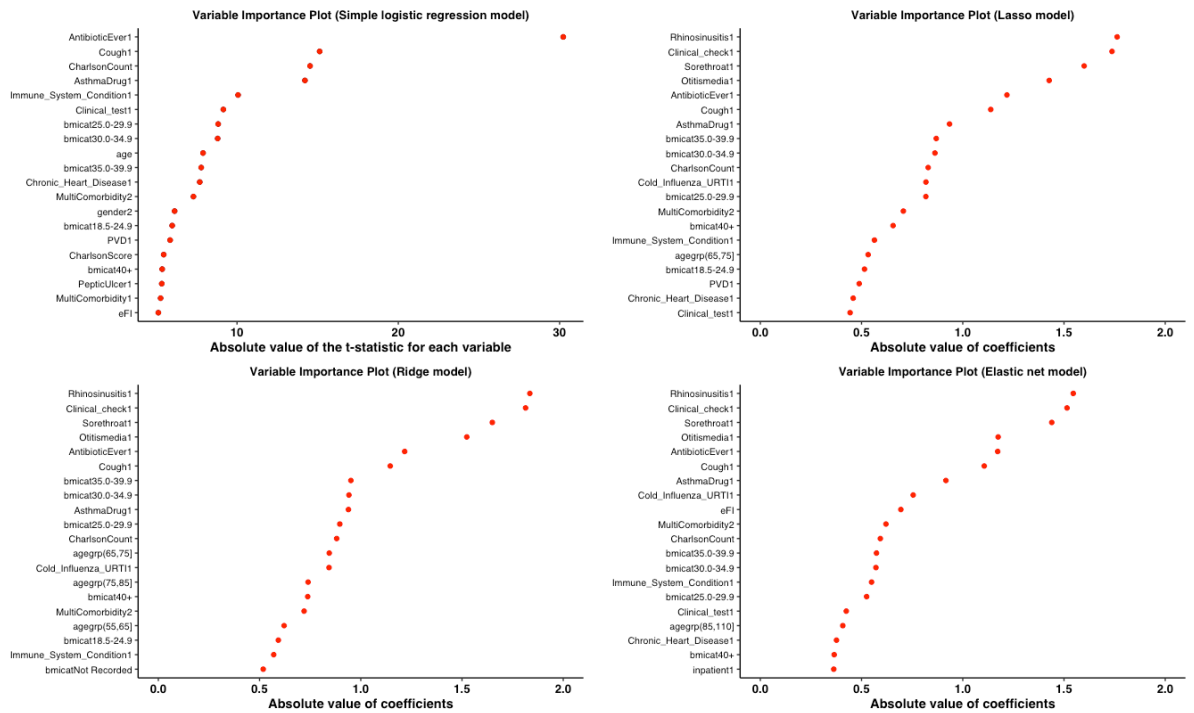


Figure A 9: Variable importance for LRTI model based on simple logistic regression and penalized regressions

Table A 21: Variable importance ranking based on top 18 variables across three penalized regression models (LRTI model)

	Lasso model	Ridge model	Elastic net model	Overall
Rhinosinusitis	20	20	20	60
Clinical check	19	19	19	57
Sore throat	18	18	18	54
Otitis media	17	17	17	51
Antibiotic ever	16	16	16	48
Cough	15	15	15	45
Asthma drug	14	12	14	40
BMI category	13	14	9	36
Charlson count	11	10	10	31
Cold/ Influenza/ URTI	10	8	13	31
Multi-comorbidity	8	5	11	24
Age group	5	9	4	18
Immune system condition	6	2	7	15
eFrailty index	0	0	12	12
Chronic heart disease	2	1	3	6
Clinical test	1	0	5	6
PVD	3	1	0	4
Inpatient	0	0	1	1

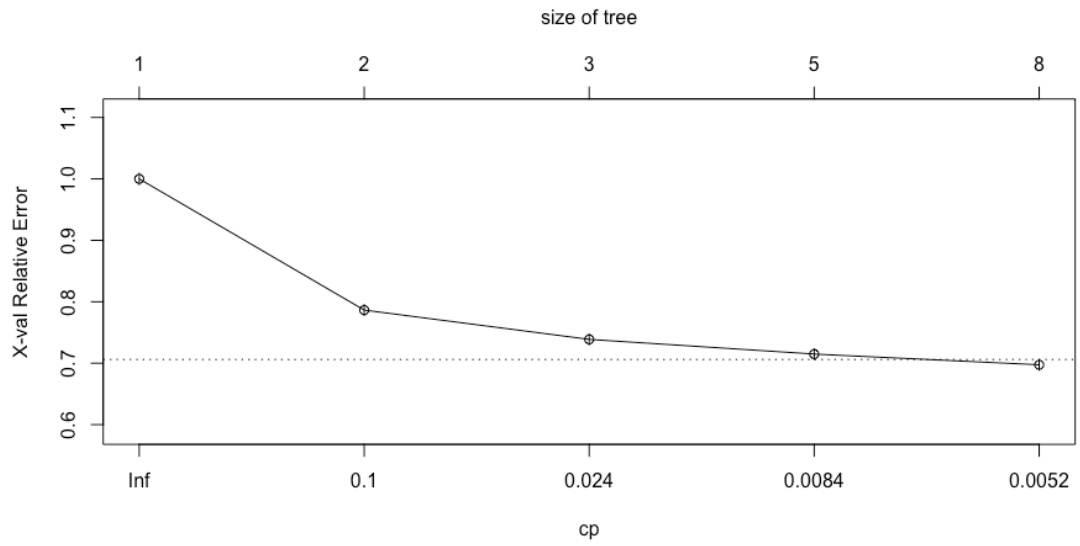
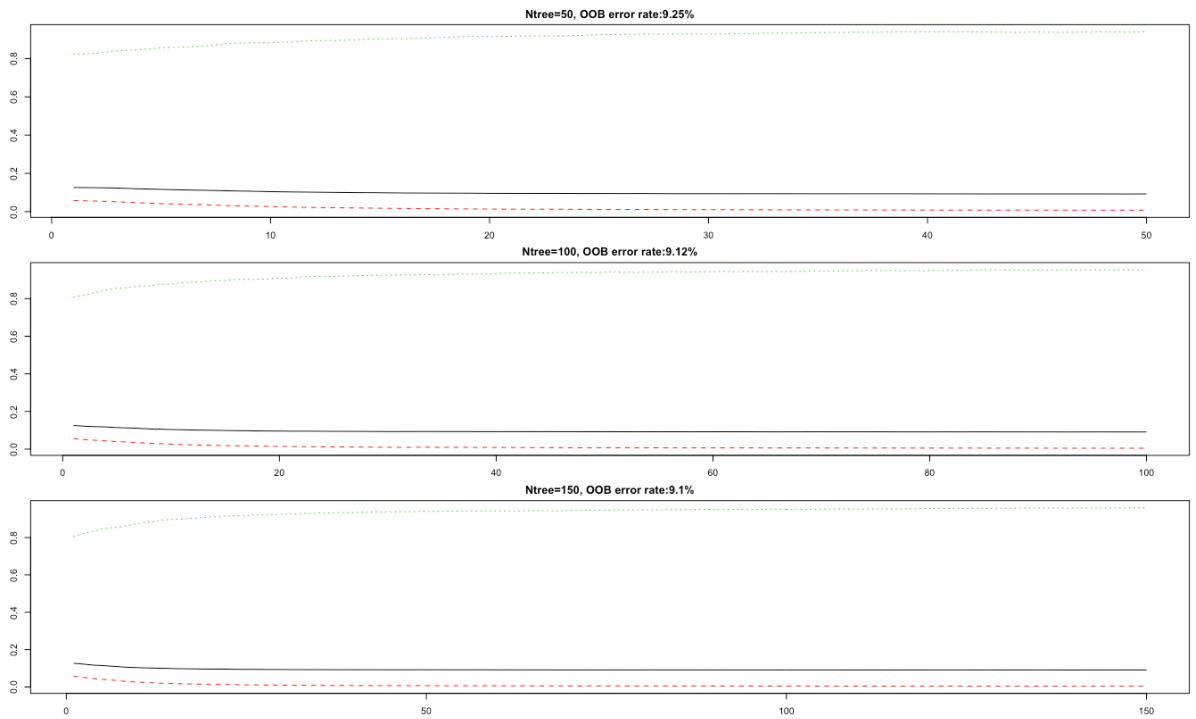


Figure A 10: Tuning parameter for CART LRTI model

Table A 22: Comparison statistics of model variables for development data and temporal validation data (LRTI model). Figures are frequencies (column percentages) except where indicated.

	Development data		Temporal validation data	
	Non-pneumonia patients (n=9,461)	Pneumonia patients (n=7,942)	Non-pneumonia patients (n=1,338)	Pneumonia patients (n=1,404)
Age group				
16-35	1457 (15.4)	562 (7.1)	207 (15.5)	87 (6.2)
36-45	1300 (13.7)	691 (8.7)	152 (11.4)	92 (6.6)
46-55	1476 (15.6)	797 (10.0)	228 (17.0)	160 (11.4)
56-65	1770 (18.7)	1168 (14.7)	259 (19.4)	183 (13.0)
66-75	1630 (17.2)	1447 (18.2)	216 (16.1)	272 (19.4)
76-85	1329 (14.0)	1873 (23.6)	195 (14.6)	331 (23.6)
86 and above	499 (5.3)	1404 (17.7)	81 (6.1)	279 (19.9)
Charlson Comorbidity Count				
0	4118 (43.5)	2306 (29.0)	589 (44.0)	342 (24.4)
1	3319 (35.1)	2464 (31.0)	448 (33.5)	432 (30.8)
2	1201 (12.7)	1517 (19.1)	155 (11.6)	292 (20.8)
3	525 (5.5)	903 (11.4)	92 (6.9)	162 (11.5)
4	194 (2.1)	458 (5.8)	37 (2.8)	105 (7.5)
5	84 (0.9)	207 (2.6)	14 (1.0)	50 (3.6)
6	20 (0.2)	87 (1.1)	3 (0.2)	21 (1.5)
Antibiotic prescription on the RTI index date	8155 (86.2)	4928 (62.0)	1125 (84.1)	902 (64.2)
Asthma drug use	1379 (14.6)	491 (6.2)	142 (10.6)	83 (5.9)
Immune system condition (Yes)	2537 (26.8)	1592 (20.0)	451 (33.7)	344 (24.5)



Green line: misclassification rate of pneumonia category

Black line: overall misclassification rate of both categories

Red line: misclassification rate of non-pneumonia category

Figure A 11: Number of trees for random forest 50, 100 and 150 for URTI model

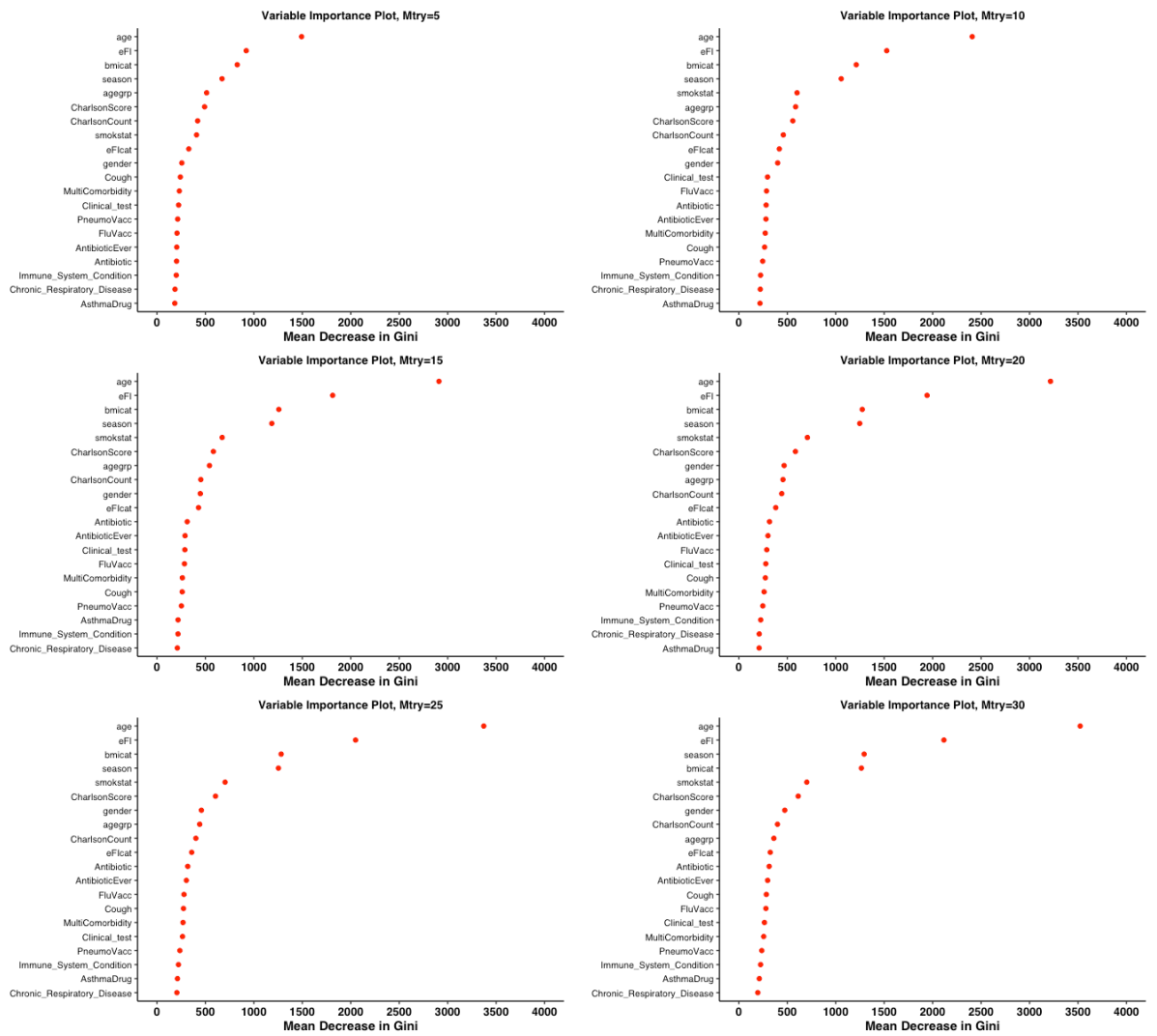


Figure A 12: Variable importance results by random forest models with Mtry 5-30 (5 increments) for URTI model

Table A 23: Variable importance ranking based on top 10 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for URTI model

	Mtry5	Mtry10	Mtry15	Mtry20	Mtry25	Mtry30	Overall
Age	10	10	10	10	10	10	60
eFrailty Index	9	9	9	9	9	9	54
BMI category	8	8	8	8	8	7	47
Season	7	7	7	7	7	8	43
Smoking Status	3	6	6	6	6	6	33
Charlson Score	5	4	5	5	5	5	29
Age group	6	5	4	3	3	2	23
Charlson Count	4	3	3	2	2	3	17
Gender	1	1	2	4	4	4	16
eFrailty category	2	2	1	1	1	1	8

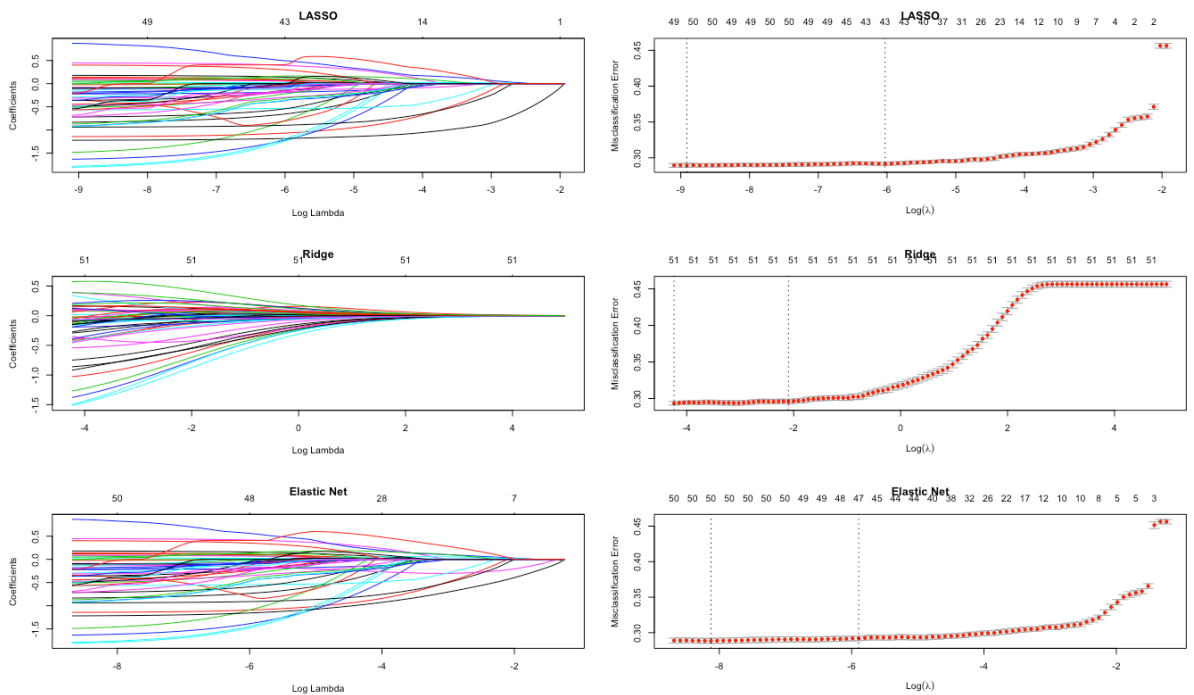


Figure A 13: Lambda for 3 penalize regressions (URTI model)

Table A 24: Lambda and alpha for elastic net (URTI model)

alpha	mim lamda	mim lamda+1se	misclassification error (mim lamda)	misclassification error (mim lamda+1se)
0	0.175888	63.06531	0.091219	0.091285
0.01	0.165276	6.306531	0.091219	0.091285
0.02	0.089787	3.153265	0.091219	0.091285
0.03	0.10124	2.102177	0.091219	0.091285
0.04	0.071184	1.576633	0.091186	0.091285
0.05	0.069113	1.261306	0.091111	0.091285
0.06	0.062003	1.051088	0.091121	0.091285
<i>0.07</i>	<i>0.057213</i>	<i>0.900933</i>	<i>0.091099</i>	<i>0.091285</i>
0.08	0.057486	0.788316	0.091165	0.091285
0.09	0.052532	0.700726	0.091121	0.091285
0.1	0.048605	0.630653	0.091143	0.091285
0.11	0.045008	0.573321	0.091143	0.091285
0.12	0.043604	0.525544	0.091154	0.091285
0.13	0.039515	0.485118	0.091121	0.091285
0.14	0.037721	0.450466	0.091154	0.091285
0.15	0.038607	0.420435	0.091121	0.091285
0.16	0.037901	0.394158	0.091111	0.091285
0.17	0.033752	0.370972	0.091176	0.091285
0.18	0.034956	0.350363	0.091121	0.091285
0.19	0.034045	0.331923	0.091099	0.091285
0.2	0.034182	0.315327	0.091132	0.091285
0.21	0.034406	0.300311	0.091111	0.091285
0.22	0.031946	0.28666	0.091132	0.091285
0.23	0.028913	0.274197	0.091143	0.091285
0.24	0.030384	0.262772	0.091111	0.091285
0.25	0.028901	0.252261	0.091111	0.091285
0.26	0.028569	0.242559	0.091111	0.091285
0.27	0.02751	0.233575	0.091132	0.091285
0.28	0.0251	0.225233	0.091121	0.091285
0.29	0.029684	0.217467	0.091132	0.091285
0.3	0.027912	0.210218	0.091154	0.091285
0.31	0.026275	0.203436	0.091132	0.091285
0.32	0.02641	0.197079	0.091121	0.091285
0.33	0.026572	0.191107	0.091132	0.091285

0.34	0.024178	0.185486	0.091132	0.091285
0.35	0.025285	0.180187	0.091121	0.091285
0.36	0.022626	0.175181	0.091121	0.091285
0.37	0.02414	0.170447	0.091121	0.091285
0.38	0.022864	0.165961	0.091154	0.091285
0.39	0.021079	0.161706	0.091154	0.091285
0.4	0.023818	0.157663	0.091165	0.091285
0.41	0.020996	0.153818	0.091165	0.091285
0.42	0.020122	0.150155	0.091154	0.091285
0.43	0.02689	0.146664	0.091197	0.091285
0.44	0.026037	0.14333	0.091154	0.091285
0.45	0.019666	0.140145	0.091143	0.091285
0.46	0.022712	0.137098	0.091154	0.091285
0.47	0.020839	0.134182	0.091154	0.091285
0.48	0.021368	0.131386	0.091176	0.091285
0.49	0.018568	0.128705	0.091143	0.091285
0.5	0.019954	0.126131	0.091165	0.091285
0.51	0.020675	0.123657	0.091143	0.091285
0.52	0.023938	0.121279	0.091197	0.091285
0.53	0.019174	0.118991	0.091143	0.091285
0.54	0.017971	0.116788	0.091132	0.091285
0.55	0.020449	0.114664	0.091165	0.091285
0.56	0.017981	0.112617	0.091165	0.091285
0.57	0.018499	0.110641	0.091165	0.091285
0.58	0.016887	0.108733	0.091143	0.091285
0.59	0.01924	0.10689	0.091165	0.091285
0.6	0.016937	0.105109	0.091165	0.091285
0.61	0.018609	0.103386	0.091154	0.091285
0.62	0.015944	0.101718	0.091165	0.091285
0.63	0.016892	0.100104	0.091154	0.091285
0.64	0.016475	0.09854	0.091154	0.091285
0.65	0.014793	0.097024	0.091165	0.091285
0.66	0.014978	0.095554	0.091154	0.091285
0.67	0.015593	0.094127	0.091132	0.091285
0.68	0.014945	0.092743	0.091154	0.091285
0.69	0.014593	0.091399	0.091143	0.091285
0.7	0.014122	0.090093	0.091154	0.091285
0.71	0.014446	0.088824	0.091165	0.091285
0.72	0.01451	0.087591	0.091154	0.091285

0.73	0.013667	0.086391	0.091165	0.091285
0.74	0.013989	0.085223	0.091154	0.091285
0.75	0.014722	0.084087	0.091154	0.091285
0.76	0.013372	0.082981	0.091154	0.091285
0.77	0.013444	0.081903	0.091165	0.091285
0.78	0.013149	0.080853	0.091176	0.091285
0.79	0.015756	0.07983	0.091176	0.091285
0.8	0.015704	0.078832	0.091186	0.091285
0.81	0.013632	0.077858	0.091132	0.091285
0.82	0.012167	0.076909	0.091176	0.091285
0.83	0.012704	0.075982	0.091165	0.091285
0.84	0.012669	0.075078	0.091154	0.091285
0.85	0.011956	0.074194	0.091132	0.091285
0.86	0.011708	0.073332	0.09111	0.091285
0.87	0.011898	0.072489	0.091186	0.091285
0.88	0.012318	0.071665	0.091154	0.091285
0.89	0.011524	0.07086	0.091165	0.091285
0.9	0.013704	0.070073	0.091197	0.091285
0.91	0.011912	0.069303	0.091186	0.091285
0.92	0.011356	0.068549	0.091186	0.091285
0.93	0.011443	0.067812	0.091154	0.091285
0.94	0.013	0.067091	0.091197	0.091285
0.95	0.010599	0.066385	0.091165	0.091285
0.96	0.010586	0.065693	0.091154	0.091285
0.97	0.010971	0.065016	0.091154	0.091285
0.98	0.011164	0.064352	0.091143	0.091285
0.99	0.010171	0.063702	0.091186	0.091285
1	0.010544	0.063065	0.091165	0.091285

Blue highlighted lambda and alpha are chosen for the elastic model

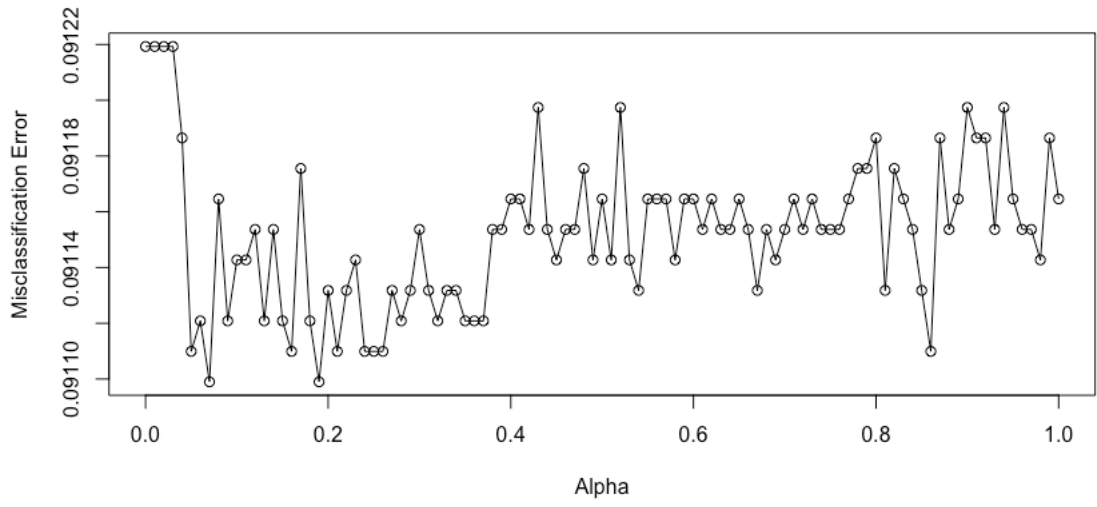


Figure A 14: Alpha for elastic net model against misclassification error (URTI model)

Table A 25: Absolute value of coefficients of three penalized regression models (URTI model)

	Lasso model		Ridge model		Elastic net model	
Variable	Importance	Sign	Importance	Sign	Importance	Sign
Age	0.02	POS	0.01	POS	0.01	POS
Age group (35,45]	0.00	NEG	0.18	POS	0.00	NEG
Age group (45,55]	0.00	NEG	0.01	POS	0.00	NEG
Age group (55,65]	0.00	NEG	0.09	POS	0.00	NEG
Age group (65,75]	0.00	NEG	0.05	POS	0.00	NEG
Age group (75,85]	0.00	NEG	0.24	POS	0.16	POS
Age group (85,110]	0.30	POS	0.62	POS	0.53	POS
Antibiotic (Yes)	0.00	NEG	0.03	POS	0.00	NEG
Antibiotic prescription in the following 30 days after initial RTI consultations	0.00	NEG	0.12	POS	0.00	POS
Asthma drug (Yes)	0.49	NEG	0.93	NEG	0.45	NEG
Healthy weight	0.00	NEG	0.03	NEG	0.03	POS
Overweight	0.00	NEG	0.28	NEG	0.03	NEG
Obese	0.00	NEG	0.30	NEG	0.03	NEG
Severe obese	0.00	NEG	0.27	NEG	0.00	NEG
Morbid obese	0.00	NEG	0.03	NEG	0.00	NEG
BMI information not recorded	0.00	NEG	0.03	NEG	0.00	NEG
Cancer (Yes)	0.00	NEG	0.07	POS	0.17	POS
Charlson Count	0.16	POS	0.26	POS	0.11	POS
Charlson Score	0.07	POS	0.10	POS	0.08	POS
Chronic heart disease (Yes)	0.00	NEG	0.12	NEG	0.00	NEG
Chronic kidney disease (Yes)	0.00	NEG	0.14	NEG	0.00	NEG
Chronic liver disease (Yes)	0.00	NEG	0.33	POS	0.03	POS
Chronic neurological condition (Yes)	0.00	NEG	0.08	POS	0.15	POS
Chronic respiratory disease (yes)	0.00	NEG	0.03	NEG	0.00	NEG

Clinical check (Yes)	0.00	NEG	0.97	NEG	0.00	NEG
Clinical test (Yes)	0.00	NEG	0.22	POS	0.06	POS
Cold/ Influenza/ URTI (Yes)	0.00	NEG	0.26	POS	0.04	POS
Cough (Yes)	0.54	POS	0.58	POS	0.38	POS
Diabetes (Yes)	0.00	NEG	0.22	NEG	0.00	NEG
eFrailty index	0.00	NEG	0.70	POS	0.57	POS
Frailty (Mild)	0.00	NEG	0.05	POS	0.00	NEG
Frailty (Moderate)	0.00	NEG	0.03	POS	0.00	POS
Frailty (Severe)	0.00	NEG	0.00	NEG	0.00	NEG
Flu vaccination (Yes)	0.00	NEG	0.07	POS	0.06	POS
Female	0.00	NEG	0.16	NEG	0.06	NEG
Hemiplegia (Yes)	0.00	NEG	0.21	POS	0.00	NEG
Immune system condition (Yes)	0.00	NEG	0.34	NEG	0.12	NEG
Hospital admission in previous year (Yes)	0.00	NEG	0.42	POS	0.05	POS
Multi-comorbidity (one comorbidity)	0.00	NEG	0.12	NEG	0.00	NEG
Multi-comorbidity (more than one comorbidity)	0.00	NEG	0.28	NEG	0.00	NEG
Otitis media (Yes)	0.05	NEG	0.97	NEG	0.36	NEG
Peptic ulcer (Yes)	0.00	NEG	0.21	NEG	0.00	NEG
Pneumococcal vaccination (Yes)	0.00	NEG	0.12	POS	0.12	POS
PVD (Yes)	0.00	NEG	0.16	NEG	0.00	NEG
Rhinosinusitis (Yes)	0.16	NEG	0.90	NEG	0.36	NEG
Season (spring)	0.00	NEG	0.06	NEG	0.00	NEG
Season (summer)	0.00	NEG	0.08	NEG	0.00	NEG
Season (winter)	0.00	NEG	0.02	NEG	0.00	NEG
Past smoker	0.00	NEG	0.18	POS	0.06	POS
Current smoker	0.00	NEG	0.16	POS	0.01	POS
Sore throat (Yes)	0.28	NEG	0.72	NEG	0.37	NEG

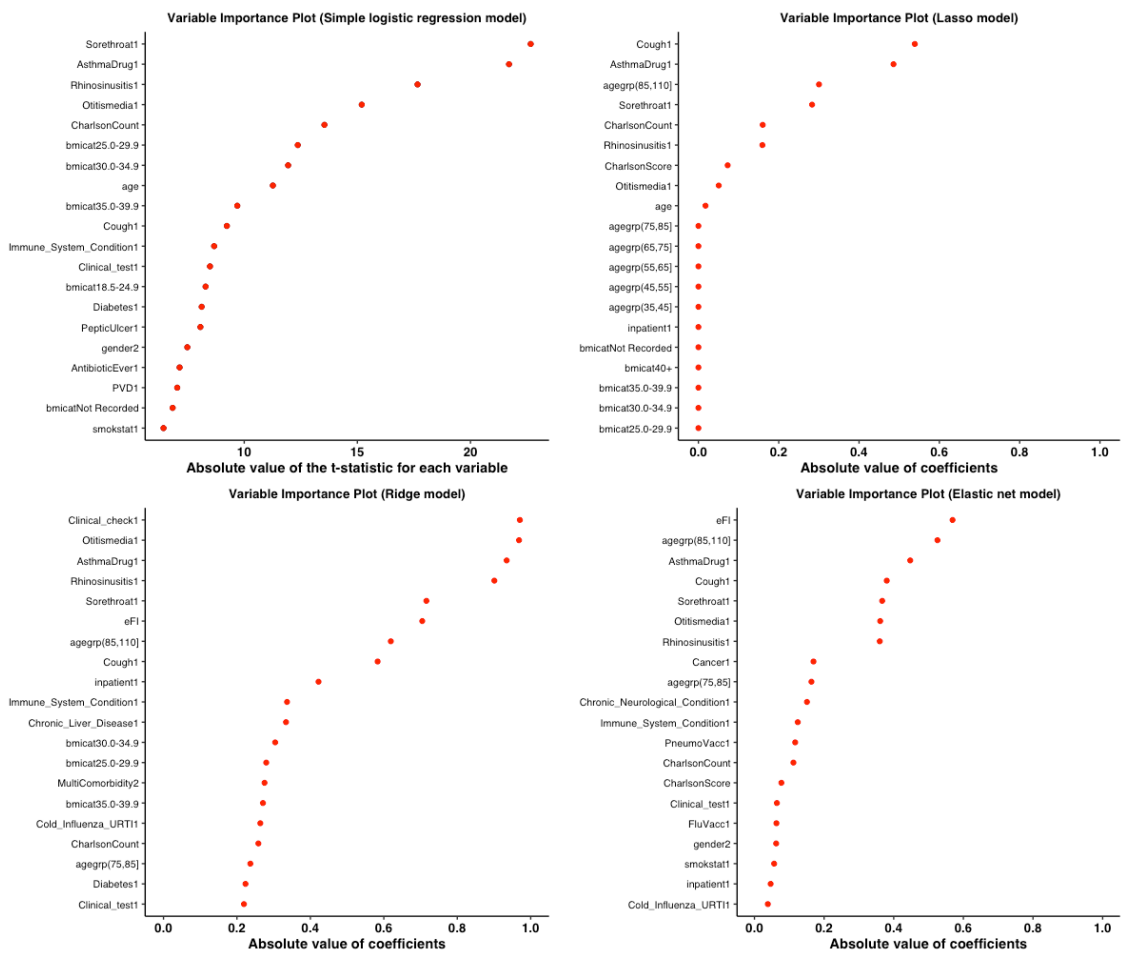


Figure A 15: Variable importance for URTI model based on simple logistic regression and penalized regressions

Table A 26: Variable importance ranking based on top 15 variables across three penalized regression models (URTI model)

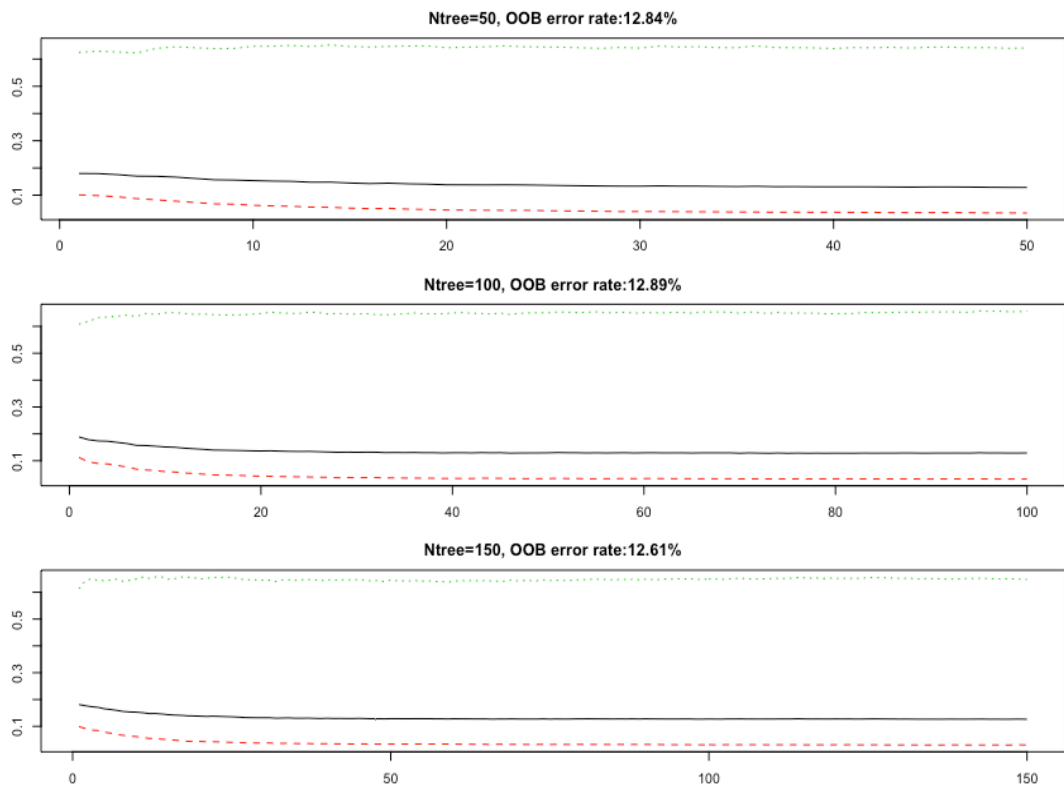
	Lasso model	Ridge model	Elastic net model	Overall
Asthma drug	9	8	8	25
Age group	8	4	9	21
Cough	10	3	7	20
Sore throat	7	6	6	19
Rhinosinusitis	4	7	4	15
eFrailty index	0	5	10	15
Otitismedia	0	9	5	14
Clinical check	0	10	0	10
Charlson Count	6	0	0	6
Charlson Score	5	0	0	5
Inpatient	2	2	0	4
Age	3	0	0	3
Cancer	0	0	3	3
BMI category	1	0	0	1
Immune system condition	0	1	0	1
Chronic neurological disease	0	0	1	1



Figure A 16: Tuning parameter for CART URTI model

Table A 27: Comparison statistics of model variables for development data and temporal validation data (URTI model). Figures are frequencies (column percentages) except where indicated.

	Development data		Temporal validation data	
	Non-pneumonia patients (n=83,092)	Pneumonia patients (n=8,347)	Non-pneumonia patients (n=16,684)	Pneumonia patients (n=1,906)
Age group				
16-35	25061 (30.2)	884 (10.6)	4852 (29.1)	176 (9.2)
36-45	14340 (17.3)	907 (10.9)	2520 (15.1)	161 (8.4)
46-55	13140 (15.8)	928 (11.1)	2707 (16.2)	205 (10.8)
56-65	12110 (14.6)	1312 (15.7)	2477 (14.8)	273 (14.3)
66-75	9977 (12.0)	1573 (18.8)	2320 (13.9)	415 (21.8)
76-85	6340 (7.6)	1676 (20.1)	1338 (8.0)	401 (21.0)
86 and above	2124 (2.6)	1067 (12.8)	470 (2.8)	275 (14.4)
Charlson Comorbidity Count				
0	2836 (34.0)	50900 (55.7)	9325 (55.9)	599 (31.4)
1	25204 (30.3)	2563 (30.7)	5126 (30.7)	574 (30.1)
2	6233 (7.5)	1447 (17.3)	1389 (8.3)	332 (17.4)
3	2341 (2.8)	831 (10.0)	553 (3.3)	208 (10.9)
4	852 (1.0)	420 (5.0)	204 (1.2)	116 (6.1)
5	280 (0.3)	164 (2.0)	63 (0.4)	54 (2.8)
6	118 (0.1)	86 (1.0)	24 (0.1)	23 (1.2)
Asthma drug use	8729 (10.5)	444 (5.3)	1505 (9.0)	81 (4.2)
Body weight				
Healthy weight (Ref)	26590 (32.0)	2888 (34.6)	5034 (30.2)	627 (32.9)
Under weight	1901 (2.3)	378 (4.5)	401 (2.4)	90 (4.7)
Overweight	23420 (28.2)	2364 (28.3)	4796 (28.7)	565 (29.6)
Obese	11372 (13.7)	1084 (13.0)	2567 (15.4)	281 (14.7)
Severe obese	4307 (5.2)	395 (4.7)	1069 (6.4)	114 (6.0)
Morbid obese	2417 (2.9)	259 (3.1)	644 (3.9)	79 (4.1)
BMI information not recorded	13085 (15.7)	979 (11.7)	2173 (13.0)	150 (7.9)



Green line: misclassification rate of pneumonia category

Black line: overall misclassification rate of both categories

Red line: misclassification rate of non-pneumonia category

Figure A 17: Number of trees for random forest 50, 100 and 150 for sensitivity analysis of full model

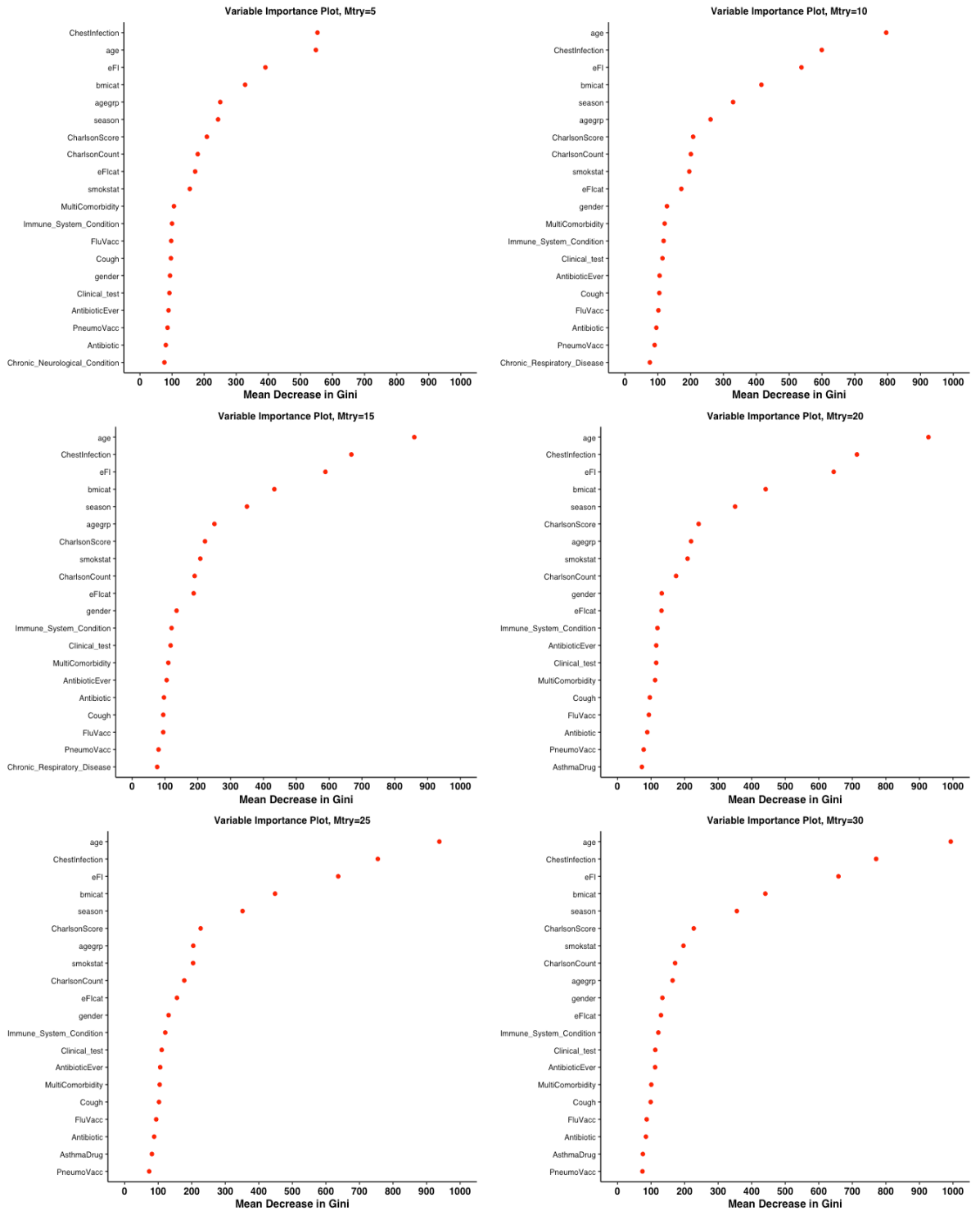


Figure A 18: Variable importance results by random forest models with Mtry 5-30 (5 increments) for sensitivity analysis of full model

Table A 28: Variable importance ranking based on top 20 variables across Mtry (5-30 in increments of 5) models (Ntree=50) for sensitivity analysis model

	Mtry5	Mtry10	Mtry15	Mtry20	Mtry25	Mtry30	Overall
Age	19	20	20	20	20	20	119
Chest infection	20	19	19	19	19	19	115
eFrailty Index	18	18	18	18	18	18	108
BMI category	17	17	17	17	17	17	102
Season	15	16	16	16	16	16	95
Charlson Score	14	14	14	15	15	15	87
Age group	16	15	15	14	14	12	86
Smoking status	11	12	13	13	13	14	76
Charlson Count	13	13	12	12	12	13	75
eFrailty category	12	11	11	10	11	10	65
Gender	6	10	10	11	10	11	58
Immune system condition	9	8	9	9	9	9	53
Multi-comorbidity	10	9	7	6	6	6	44
Clinical test	5	7	8	7	8	8	43
Antibiotic Ever	4	6	6	8	7	7	38
Cough	7	5	4	5	5	5	31
Flu vaccination	8	4	3	4	4	4	27
Antibiotic	2	3	5	3	3	3	19
Pneumococcal Vaccination	3	2	2	2	1	1	11
Asthma drug	0	0	0	1	2	2	5

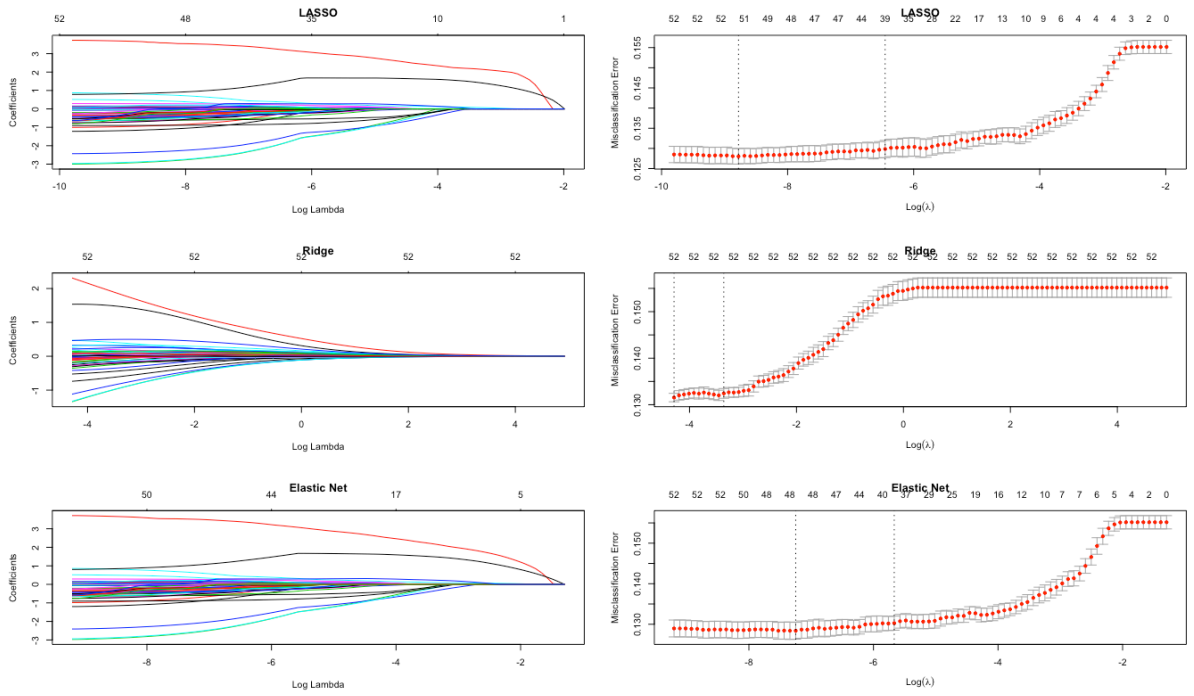


Figure A 19: Lambda for 3 penalize regressions (sensitivity analysis model of full model)

Table A 29: Lambda and alpha for elastic net (sensitivity analysis of full model)

alpha	mim lamda	mim lamda+1se	misclassification error (mim lamda)	misclassification error (mim lamda+1se)
0	0.028133	0.054138	0.13168	0.133696
0.01	0.002318	0.031208	0.129711	0.131868
0.02	0.002757	0.004929	0.129383	0.130227
0.03	0.002446	0.009313	0.128961	0.130696
0.04	0.002353	0.010193	0.129524	0.131352
0.05	0.002483	0.00886	0.129102	0.131071
0.06	0.002728	0.031688	0.129102	0.132008
0.07	0.001647	0.006156	0.129336	0.131211
0.08	0.001954	0.012933	0.129196	0.131868
0.09	0.001485	0.022742	0.12868	0.131774
0.1	0.001439	0.006289	0.128586	0.130321
0.11	0.001344	0.005262	0.128399	0.130321
0.12	0.001428	0.00613	0.129149	0.130743
0.13	0.001419	0.013838	0.128821	0.131211
0.14	0.001191	0.007123	0.128071	0.130789
0.15	0.001507	0.007494	0.129102	0.130321
0.16	0.001426	0.010937	0.12868	0.130977
0.17	0.001405	0.016935	0.128867	0.131821
0.18	0.001054	0.011476	0.128867	0.131258
0.19	0.001085	0.010383	0.129008	0.130602
0.2	0.001476	0.007343	0.128867	0.130649
0.21	0.00138	0.006866	0.128586	0.130883
0.22	0.001096	0.001995	0.128821	0.129758
0.23	0.001602	0.013853	0.129008	0.13154
0.24	0.001101	0.010446	0.129102	0.130977
0.25	0.001203	0.014501	0.129008	0.131493
0.26	0.000877	0.009643	0.128211	0.130743
0.27	0.000997	0.016907	0.128586	0.132055
0.28	0.00098	0.01331	0.129336	0.132008
0.29	0.001472	0.003771	0.128258	0.129711
0.3	0.000898	0.008054	0.129196	0.130555
0.31	0.001054	0.012359	0.129102	0.131727
0.32	0.00149	0.010523	0.128961	0.131258
0.33	0.00138	0.009052	0.128727	0.131211

0.34	0.000868	0.006189	0.128727	0.130602
0.35	0.000892	0.004082	0.128821	0.130086
0.36	0.001092	0.008932	0.128961	0.130461
0.37	0.000908	0.01316	0.128399	0.13168
0.38	0.000821	0.002648	0.128352	0.129664
0.39	0.001257	0.006548	0.128727	0.130508
0.4	0.00126	0.008496	0.129055	0.130743
0.41	0.000866	0.006767	0.128399	0.130883
0.42	0.000877	0.005914	0.128727	0.130461
0.43	0.00065	0.003544	0.128258	0.129899
0.44	0.001026	0.006852	0.128727	0.130743
0.45	0.000644	0.005781	0.128774	0.130602
0.46	0.000697	0.005063	0.128727	0.130602
0.47	0.000806	0.012343	0.128867	0.132336
0.48	0.000615	0.008834	0.128586	0.131118
0.49	0.000679	0.007263	0.128821	0.131071
0.5	0.000654	0.004968	0.128399	0.130321
0.51	0.000606	0.007795	0.128774	0.13154
0.52	0.001073	0.006907	0.128539	0.130602
0.53	0.0006	0.004394	0.128586	0.130227
0.54	0.000594	0.005846	0.129008	0.130508
0.55	0.000714	0.004643	0.128446	0.130133
0.56	0.000562	0.007502	0.129336	0.131446
0.57	0.000646	0.004564	0.128774	0.130321
0.58	0.001064	0.006136	0.128633	0.130883
0.59	0.000758	0.00497	0.128164	0.130133
0.6	0.000637	0.007002	0.128492	0.13168
0.61	0.000668	0.008514	0.128961	0.132149
0.62	0.000715	0.005381	0.129008	0.130508
0.63	0.000716	0.007381	0.12868	0.131211
0.64	0.000649	0.005213	0.128492	0.130743
0.65	0.00072	0.008762	0.129008	0.131915
0.66	0.000635	0.005493	0.128774	0.130696
0.67	0.000643	0.004934	0.128492	0.130461
0.68	0.000836	0.004817	0.128821	0.130836
0.69	0.00066	0.006089	0.128914	0.131305
0.7	0.000616	0.004594	0.129008	0.130883
0.71	0.000691	0.003981	0.12793	0.13018
0.72	0.000797	0.002817	0.128305	0.129993

0.73	0.000601	0.006253	0.128914	0.131493
0.74	0.000663	0.005891	0.128492	0.131305
0.75	0.000602	0.004616	0.128821	0.130977
0.76	0.000588	0.006586	0.129102	0.131633
0.77	0.000711	0.005406	0.129008	0.131493
0.78	0.0006	0.005852	0.128305	0.131118
0.79	0.0007	0.005417	0.128539	0.131071
0.8	0.001169	0.005759	0.129524	0.13168
0.81	0.001508	0.014701	0.129664	0.132946
0.82	0.000598	0.007616	0.128867	0.132008
0.83	0.000619	0.001753	0.128774	0.130321
0.84	0.000676	0.001363	0.128914	0.130133
0.85	0.0007	0.004591	0.128821	0.131071
0.86	0.000787	0.006152	0.128821	0.131493
0.87	0.000716	0.004612	0.129289	0.130555
0.88	0.001198	0.004476	0.128867	0.13093
0.89	0.00072	0.006168	0.129102	0.13168
0.9	0.000655	0.002338	0.129055	0.130274
0.91	0.00073	0.004491	0.128961	0.130555
0.92	0.00075	0.005441	0.129008	0.13168
0.93	0.000702	0.005093	0.129196	0.131352
0.94	0.000707	0.004767	0.128633	0.130696
0.95	0.000693	0.004422	0.129243	0.131493
0.96	0.000661	0.003846	0.128961	0.131024
0.97	0.000654	0.003737	0.128821	0.130696
0.98	0.000697	0.004923	0.128867	0.131352
0.99	0.000612	0.003628	0.129149	0.131024
1	0.000617	0.002826	0.128492	0.130321

Blue highlighted lambda and alpha are chosen for the elastic model

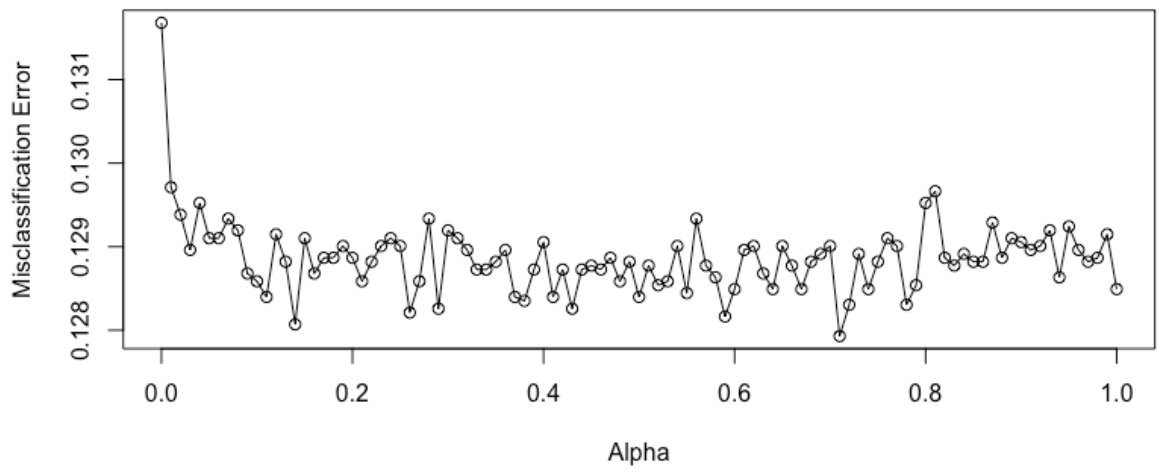


Figure A 20: Alpha for elastic net model against misclassification error (sensitivity analysis of full model)

Table A 30: Absolute value of coefficients of three penalized regression models (sensitivity analysis of full model)

	Lasso model		Ridge model		Elastic net model	
Variable	Importance	Sign	Importance	Sign	Importance	Sign
Age	0.02995476	POS	0.02941424	POS	0.01826777	POS
Age group (35,45]	0.09415568	POS	0.10100312	POS	0.22036003	POS
Age group (45,55]	0.22518157	NEG	0.2177951	NEG	0.02556904	POS
Age group (55,65]	0.49431343	NEG	0.48227576	NEG	0.0963749	NEG
Age group (65,75]	0.59498546	NEG	0.57970724	NEG	0.07383369	NEG
Age group (75,85]	0.73670173	NEG	0.71601903	NEG	0.09870252	NEG
Age group (85,110]	0.55296398	NEG	0.52734249	NEG	0.17826426	POS
Antibiotic (Yes)	0.05236503	NEG	0.06754535	NEG	0	NEG
Antibiotic prescription in the following 30 days after initial RTI consultations	0.07590919	POS	0.08998655	POS	0.01954564	POS
Asthma drug (Yes)	0.90276116	NEG	0.90123754	NEG	0.87641046	NEG
Healthy weight	0.40632628	NEG	0.41873195	NEG	0.16675671	NEG
Overweight	0.63628078	NEG	0.64794475	NEG	0.394563	NEG
Obese	0.72779972	NEG	0.73905069	NEG	0.47950306	NEG
Severe obese	0.74258474	NEG	0.75684039	NEG	0.49094253	NEG
Morbid obese	0.50291967	NEG	0.51863441	NEG	0.24839324	NEG
BMI information not recorded	0.20196261	NEG	0.21882077	NEG	0	NEG
Cancer (Yes)	0.02470437	POS	0.0096775	POS	0	NEG
Charlson Count	0.8641132	POS	0.81429869	POS	0.60903827	POS
Charlson Score	0.22591801	NEG	0.19434034	NEG	0.08892039	NEG
Chest Infection (Yes)	0.80927988	POS	0.87422914	POS	1.06343828	POS
Chronic heart disease (Yes)	0.33209217	NEG	0.32217147	NEG	0.24208954	NEG
Chronic kidney disease (Yes)	0.20735602	NEG	0.22194754	NEG	0.21540679	NEG
Chronic liver disease (Yes)	0.05686239	POS	0.06453583	POS	0.08078093	POS

Chronic neurological condition (Yes)	0.27204981	NEG	0.25840891	NEG	0.15248804	NEG
Chronic respiratory disease (yes)	0.26628993	NEG	0.25554914	NEG	0.17201866	NEG
Clinical check (Yes)	0.4988551	POS	0.54031677	POS	0.31788492	POS
Clinical test (Yes)	0.28868622	POS	0.28723852	POS	0.26700974	POS
Cold/ Influenza/ URTI (Yes)	1.19153614	NEG	1.12307684	NEG	0.89747285	NEG
Cough (Yes)	0.97740965	NEG	0.9095084	NEG	0.69175265	NEG
Diabetes (Yes)	0.53860467	NEG	0.52898281	NEG	0.44022663	NEG
eFrailty index	3.68450489	POS	3.64882563	POS	3.48432949	POS
Frailty (Mild)	0.07884627	POS	0.08627712	POS	0.07221902	POS
Frailty (Moderate)	0.16031738	POS	0.17189769	POS	0.16362032	POS
Frailty (Severe)	0.04101509	NEG	0.02511115	NEG	0	NEG
Flu vaccination (Yes)	0.07552045	POS	0.07804079	POS	0.0614592	POS
Female	0.23181195	NEG	0.23320825	NEG	0.21067776	NEG
Hemiplegia (Yes)	0.35579161	NEG	0.37482161	NEG	0.31081294	NEG
Immune system condition (Yes)	0.62492637	NEG	0.6241592	NEG	0.59656774	NEG
Hospital admission in previous year (Yes)	0.43135151	NEG	0.44526866	NEG	0.38757609	NEG
Multi-comorbidity (one comorbidity)	0.2322179	NEG	0.23012335	NEG	0.18546577	NEG
Multi-comorbidity (more than one comorbidity)	0.62460743	NEG	0.6171946	NEG	0.56550926	NEG
Otitis media (Yes)	2.96467392	NEG	2.86715521	NEG	2.55035002	NEG
Peptic ulcer (Yes)	0.5261811	NEG	0.51207296	NEG	0.39210211	NEG
Pneumococcal vaccination (Yes)	0.06708876	POS	0.07087085	POS	0.04978911	POS
PVD (Yes)	0.5047018	NEG	0.48928184	NEG	0.36314474	NEG
Rhinosinusitis (Yes)	2.92803986	NEG	2.82784822	NEG	2.5228787	NEG
Season (spring)	0.05559457	NEG	0.05908393	NEG	0.03635594	NEG
Season (summer)	0.01992114	POS	0.01957233	POS	0.01776787	POS
Season (winter)	0.06963424	NEG	0.07266139	NEG	0.05028616	NEG
Past smoker	0.11284871	POS	0.11491821	POS	0.09438234	POS

Current smoker	0.0010291	POS	0.0036448	POS	0	NEG
Sore throat (Yes)	2.4030107	NEG	2.31778503	NEG	2.0787913	NEG

Table A 31: Variable importance ranking based on top 20 variables across three penalized regression models (sensitivity analysis of full model)

	Lasso model	Ridge model	Elastic net model	Overall
eFI	20	20	20	60
Otitismedia	19	19	19	57
Rhinosinusitis	18	18	18	54
Sorethroat	17	17	17	51
Cold_Influenza_URT11	16	16	15	47
Cough	15	15	13	43
AsthmaDrug	14	13	14	41
Chest Infection	12	11	16	39
Charlson Count	13	12	12	37
BMI category	11	10	9	30
Immune System Condition	9	9	11	29
Multi-Comorbidity	8	8	10	26
Age group	10	14	1	25
Diabetes	7	7	8	22
PepticUlcer	6	5	7	18
PVD	5	4	5	14
Clinial check	4	6	4	14
Admission during previous year	3	3	6	12
Hemiplegia	2	2	3	7
Chronic heart disease	1	1	0	2

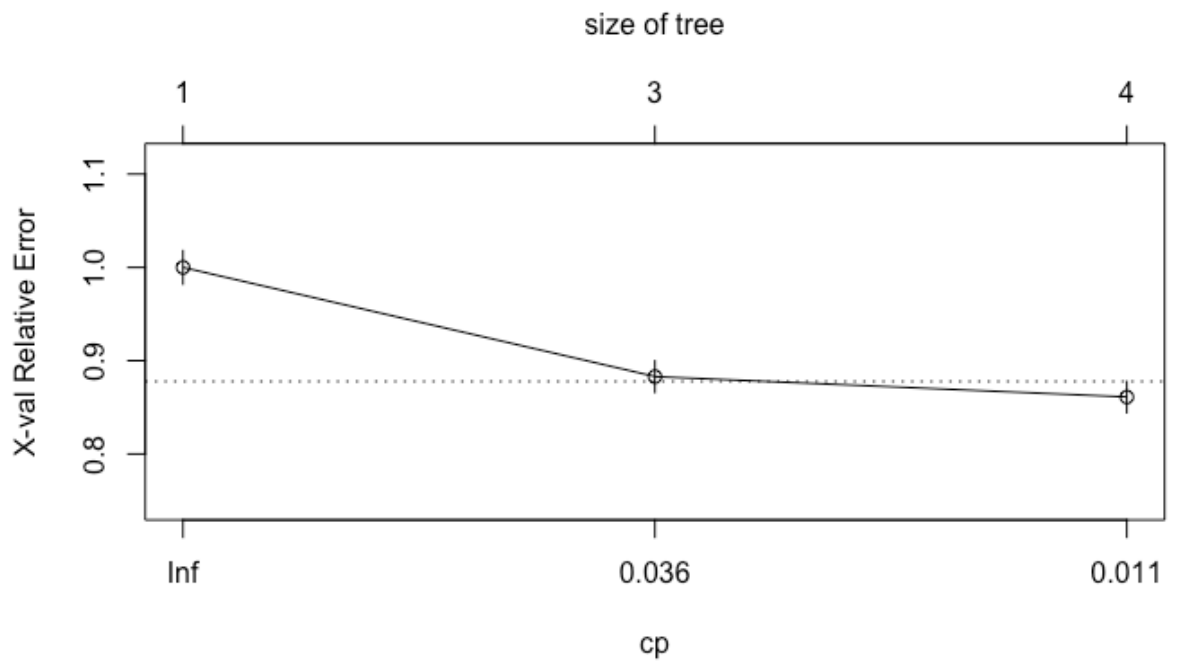


Figure A 21: Tuning parameter for CART model (sensitivity analysis of full model)

Table A 32: Comparison statistics of model variables for development data and temporal validation data (sensitivity analysis of full model). Figures are frequencies (column percentages) except where indicated.

	Development data		Temporal validation data	
	Non-pneumonia patients (n=18,022)	Pneumonia patients (n=3,310)	Non-pneumonia patients (n=3,376)	Pneumonia patients (n=629)
Age group				
16-35	5059 (28.1)	263 (7.9)	914 (27.1)	36 (5.7)
36-45	2672 (14.8)	253 (7.6)	524 (15.6)	54 (8.6)
46-55	2935 (16.3)	365 (11.0)	551 (16.4)	79 (12.6)
56-65	2736 (15.2)	456 (13.8)	545 (16.2)	94 (14.9)
66-75	2536 (14.1)	687 (20.8)	475 (14.1)	129 (20.5)
76-85	1533 (8.5)	732 (22.1)	271 (8.0)	147 (23.4)
86 and above	551 (3.1)	554 (16.7)	87 (2.6)	90 (14.3)
Immune system condition				
Antibiotic prescription				
Chest infection	3230 (17.9)	705 (21.3)	641 (19.0)	139 (22.1)
	8708 (48.3)	1863 (56.3)	1572 (46.7)	355 (56.4)
	1,338 (7.4)	1,404 (42.4)	204 (6.1)	241 (38.3)

Appendix G: TRIPOD Checklist for reporting prediction modelling study

Table A 33: TRIPOD checklist for prediction model development and validation

Section/Topic	Item		Checklist Item	Section
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Chapter 7
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Chapter 7 and 8
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	Chapter 2 and 7
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	Section 7.2
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	Section 7.2.3
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Section 7.2.2
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	Section 7.2.2
	5b	D;V	Describe eligibility criteria for participants.	Section 7.2.3
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	Section 7.2.3
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Section 7.2.4
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	Section 7.2.3
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Section 7.1
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	Section 7.2.5
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	Section 7.2.5
	10c	V	For validation, describe how the predictions were calculated.	Section 7.2.6
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	Section 7.2.6
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Section 8.1.4
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Section 7.2.3
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Section 8.1.1

	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Section 8.1.3-8.1.5
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	Section 8.1.1-8.1.5
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	Section 8.1.1-8.1.5
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Section 8.1.1-8.1.5
	15b	D	Explain how to use the prediction model.	Section 8.2-8.3
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	Section 8.1.4-5
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	Section 8.1.5
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	Section 8.2
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	Section 8.1.4-5
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	Section 8.2
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	Section 8.3
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Appendix
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	Acknowledgement

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Appendix H: Publications and outputs during thesis

First authored publications:

SUN, X. & GULLIFORD, M. C. 2019. Reducing antibiotic prescribing in primary care in England from 2014 to 2017: population-based cohort study. *BMJ open*, 9, e023989.

SUN, X., DOUIRI, A. & GULLIFORD, M. 2019. Pneumonia incidence trends in UK primary care from 2002 to 2017: population-based cohort study. *Epidemiology & Infection*, 147.

Invited editorial publication:

SUN, X., RAHMAN, N. M. & DOUIRI, A. 2020. Treating tuberculosis in low-resource settings: practice pragmatically. *Thorax*, 75, 363-363.

Joint first authored publication:

LIANG, S.-F., SUN, X., GULLIFORD, M. & CURCIN, V. Inclusion and Exclusion of Medical Codes for Primary Care Data Extraction. 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018. IEEE, 394-395.

Co-authored publications:

GULLIFORD, M. C., CHARLTON, J., WINTER, J. R., SUN, X., REZEL-POTTS, E., BUNCE, C., FOX, R., LITTLE, P., HAY, A. D. & MOORE, M. V. 2020. Probability of sepsis after infection consultations in primary care in the United Kingdom in 2002–2017: Population-based cohort study and decision analytic model. *PLoS Medicine*, 17, e1003202.

GULLIFORD, M. C., SUN, X., ANJUMAN, T., YELLAND, E. & MURRAY-THOMAS, T. 2020. Comparison of antibiotic prescribing records in two UK primary care electronic health record systems: cohort study using CPRD GOLD and CPRD Aurum databases. *BMJ open*, 10, e038767.

GULLIFORD, M. C., SUN, X., CHARLTON, J., WINTER, J. R., BUNCE, C., BOIKO, O., FOX, R., LITTLE, P., MOORE, M. & HAY, A. D. 2020. Serious bacterial infections and antibiotic prescribing in primary care: cohort study using electronic health records in the UK. *BMJ open*, 10, e036975.

Conference presentations:

SUN, X. & GULLIFORD, M. C. 2018. Reducing antibiotic prescribing in primary care in England from 2014 to 2017: population-based cohort study. *Public Health England Annual Conference 2018*

LIANG, S.-F., SUN, X., GULLIFORD, M. & CURCIN, V. Inclusion and Exclusion of Medical Codes for Primary Care Data Extraction. *IEEE International Conference on Healthcare Informatics (ICHI) 2018*

SUN, X., DOUIRI, A. & GULLIFORD, M. 2019. Pneumonia incidence trends in UK primary care from 2002 to 2017: population-based cohort study. *European Respiratory Society (ERS) Congress 2019*

BMJ Open Reducing antibiotic prescribing in primary care in England from 2014 to 2017: population-based cohort study

Xiaohui Sun,¹ Martin C Gulliford²

To cite: Sun X, Gulliford MC. Reducing antibiotic prescribing in primary care in England from 2014 to 2017: population-based cohort study. *BMJ Open* 2019;9:e023989. doi:10.1136/bmjopen-2018-023989

► Prepublication history and additional material for this paper are available online. To view please visit the journal (<http://dx.doi.org/10.1136/bmjopen-2018-023989>).

Received 10 May 2018
Revised 14 May 2019
Accepted 15 May 2019



© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY. Published by BMJ.

School of Population and Environmental Health Sciences, King's College London, London, UK

Correspondence to
Xiaohui Sun;
xiaohui.sun@kcl.ac.uk

ABSTRACT

Objective To analyse individual-patient electronic health records to evaluate changes in antibiotic (AB) prescribing in England for different age groups, for male and female subjects, and by prescribing indications from 2014 to 2017.

Methods Data were analysed for 102 general practices in England that contributed data to the UK Clinical Practice Research Datalink (CPRD) from 2014 to 2017. Prescriptions for all ABs and for broad-spectrum β -lactam ABs were evaluated. Relative rate reductions (RRR) were estimated from a random-effects Poisson model, adjusting for age, gender, and general practice.

Results Total AB prescribing declined from 608 prescriptions per 1000 person-years in 2014 to 489 per 1000 person-years in 2017; RRR 6.9% (95% CI 6.6% to 7.1%) per year. Broad-spectrum β -lactam AB prescribing decreased from 221 per 1000 person-years in 2014 to 163 per 1000 person-years in 2017; RRR 9.3% (9.0% to 9.6%) per year. Declines in AB prescribing were similar for men and women but the rate of decline was lower over the age of 55 years than for younger patients. All AB prescribing declined by 9.8% (9.6% to 10.1%) per year for respiratory infections, 5.7% (5.2% to 6.2%) for genitourinary infections, but by 3.8% (3.1% to 4.5%) for no recorded indication. Overall, 38.8% of AB prescriptions were associated with codes that did not suggest specific clinical conditions, and 15.3% of AB prescriptions had no medical codes recorded.

Conclusion Antibiotic prescribing has reduced and become more selective but substantial unnecessary AB use may persist. Improving the quality of diagnostic coding for AB use will help to support antimicrobial stewardship efforts.

INTRODUCTION

Antimicrobial resistance is a growing concern worldwide.^{1,2} Many disease-causing pathogens have now developed resistance to antimicrobial drugs.³ The pathways to high rates of antibiotic (AB) resistance at population level are complex but excessive AB use as medical care is often a proximal cause of AB resistance,^{4,5} especially in communities.^{6–8} Consequently, there are increasing calls for more carefully considered use of ABs in order to conserve the therapeutic potential of these drugs.⁹ This

Strengths and limitations of this study

- The study findings are derived from analysis of electronic health records data for more than 100 general practices in England that continuously contributed to the CPRD dataset over the 4-year study period.
- Comprehensive data for all antibiotic prescriptions and consultations at general practice surgeries were analysed.
- Antibiotic prescriptions issued outside general practices in out-of-hours settings were not captured.
- Antibiotic prescriptions may not always have been dispensed or taken by patients.

is particularly relevant in primary care, where more than 70% of all ABs are prescribed.^{10,11} Inappropriate AB prescribing is known to be widespread in primary care.¹² Based on international comparisons, with both low¹³ and high¹⁴ AB prescribing being observed across Europe, without comparable variation in safety outcomes such as bacterial infections, it appears that a substantial reduction of present AB prescribing in primary care might be safe and feasible.

To deal with these concerns, aggregated data for AB prescribing are now being used for health service management. A contractual financial incentive, known as a 'quality premium', has been introduced into the English NHS for meeting indicative targets for year-on-year reductions in inappropriate AB use across all indications.¹⁵ The English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR)¹⁰ analysed aggregated prescribing data and found that general practice AB prescriptions decreased by 13% between 2012 and 2016. Analysis of data for individual patients offers an opportunity for more detailed understanding of this decreasing trend. Dolk *et al*¹⁶ analysed data from The Health Improvement Network (THIN) database from 2013 to 2015. They drew attention to limitations of primary care records as a data source,

including the high proportion of AB prescriptions for which no 'clinical justification' was recorded.

The purpose of this study is to update data for AB prescribing trends in English general practices from 2014 to 2017. The analyses specifically aimed to provide estimates for the decline in AB use separately for male and female subjects and for people of different ages. We also aimed to evaluate which prescribing indications were most associated with reduced prescribing. We compared changes in all AB prescribing with changes in prescribing of broad-spectrum β -lactam ABs. Finally, we aimed to compare reductions in prescribing of individual major classes of ABs to provide complementary information.

METHODS

Data source

The UK Clinical Practice Research Datalink (CPRD)¹⁷ was used as the data source for the study. This is a prospectively collected primary care database including data from approximately 7% of UK general practices. The total number of patients ever registered in CPRD is about 11 million, but the registered population has varied over time, and by 2017 there were approximately 2.5 million active UK patients. In the UK, more than 98% of the population are registered with a general practice and registrations are often maintained over many years. The CPRD is considered to be representative of the UK population.¹⁷ Data collected in the CPRD are of high quality and include all medical diagnoses recorded at consultations and referrals, as well as all drug prescriptions issued by general practices.¹⁸ For this study we included data from CPRD general practices in England, which participated in the data linkage scheme, and consistently contributed data in all years from 2014 to 2017. During this period the total number of general practices in the UK contributing to CPRD declined from 491 in 2014 to 285 in 2017. The number of CPRD general practices in England declined from 329 to 133, while the number participating in the data linkage scheme declined from 257 to 102 (online supplementary table 1). Individual participant data were included from 1 January 2014 or the start of the patient's CPRD record, whichever was the latest, to the 31 December 2017 or the end of the patient's CPRD record, whichever was the earliest. Data were obtained from the February 2018 release of CPRD. For practices that ended CPRD data collection during 2017, an equivalent end-of-year-date was also adopted for earlier years, because of the marked seasonality in AB use.

Main measures

For each year of study, we calculated the person-time contributed by each patient between 1 January of the year, or start of registration if this was later, to 31 December of the year, or end of registration or date of death, if these were earlier. Person-time was employed as the denominator for rates. Prescriptions for ABs were identified using product codes for all AB drug classes included in

section 5.1 of the *British National Formulary* (BNF) except anti-tuberculous, anti-lepromatous agents, and methenamine, which were excluded.¹⁹ The BNF groups ABs into the following categories: penicillins, cephalosporins (including carbapenems), tetracyclines, aminoglycosides, macrolides, clindamycin, sulfonamides (including combinations with trimethoprim), metronidazole and tinidazole, quinolones, drugs for urinary tract infection (nitrofurantoin), and other ABs.

We analysed broad-spectrum β -lactam ABs as a separate group, including the BNF category of 'broad-spectrum penicillins'¹⁹ and cephalosporins. The category of 'broad-spectrum penicillins' includes ampicillin and amoxicillin and combinations with clavulanic acid or flucloxacillin. Carbapenems, which are only rarely used in primary care, were combined with cephalosporins for these analyses. Clinical indications for AB prescription were grouped into categories based on Read medical codes recorded into patients' clinical and referral records on the same date as the AB prescription, including 'respiratory conditions', 'genitourinary conditions', 'skin' conditions, 'eye' conditions, or no codes recorded (online supplementary tables 2 to 5). All other codes were grouped into a single category of 'other and non-specific codes'. The most frequently used codes in this category are shown in table 1 and included 'telephone encounter', 'patient reviewed', and 'telephone triage encounter'. Since specific coded indications for AB therapy were uncommon in this category, it is subsequently referred to as 'non-specific'. We analysed the prescription sequence variable to determine whether each prescription was the first in a sequence or whether it was a repeat prescription; the former were coded as 'acute' prescriptions and the latter were coded as 'repeat' prescriptions.

Statistical analysis

Antibiotic prescriptions for all ABs and broad-spectrum β -lactam ABs were enumerated by year. AB prescriptions of the same type on the same date were considered as a single event. Age was included as a continuous covariate but was also analysed in subgroups from 0 to 4 years, then 10-year age groups up to ≥ 85 years. Read codes recorded on the same date as an AB prescription were analysed according to indication. The primary indication on each date was allocated by giving priority to indications in the following sequence: respiratory, genitourinary, skin, and eye. We estimated AB prescription rates per 1000 person-years, and proportions of registered patients with ABs prescribed in a year in relation to age group, gender, study year, and main indication. In order to estimate annual changes in AB prescribing, we fitted it to hierarchical generalised linear Poisson models using the 'hglm' package²⁰ in the R programme. The dependent variable was a count of AB prescriptions (either all AB prescriptions or broad-spectrum β -lactam AB prescriptions). Predictors were calendar year, gender, and age, including quadratic and cubic terms to allow for non-linear effects of age. Calendar year was included as a linear predictor based on inspection of

**Table 1** Thirty most frequently used Read codes for 'other and non-specific' antibiotic prescribing indications.

Read code	Read term	Number of events*
9N31.00	Telephone encounter	51 504
6A...00	Patient reviewed	32 470
9N3A.00	Telephone triage encounter	26 900
246...00	O/E - blood pressure reading	25 502
242...00	O/E - pulse rate	15 918
9Z...00	Administration NOS	9 278
22A...00	O/E - weight	8 937
8CB...00	Had a chat to patient	8 191
9N1C.11	Home visit	7 813
1371	Never smoked tobacco	6 065
9...00	Administration	5 748
8CAL.00	Smoking cessation advice	5 664
8B3H.00	Medication requested	5 661
137S.00	Ex-smoker	4 642
137P.00	Cigarette smoker	4 565
9N3G.00	SMS text message sent to patient	3 990
8B3S.00	Medication review	3 891
8CA...00	Patient given advice	3 838
246...11	O/E - BP reading	3 810
9N4...00	Failed encounter	3 514
661M.00	Clinical management plan agreed	3 305
9N58.00	Emergency appointment	2 930
1...00	History/symptoms	2 827
212...00	Patient examined	2 691
81H...00	Dressing of wound	2 543
9Na...00	Consultation	2 381
14L...00	H/O: drug allergy	2 277
1969	Abdominal pain	2 102
9N32.00	Third-party encounter	1 948
679...11	Advice to patient - subject	1 939

*Multiple codes per date were analysed. BP, blood pressure; H/O, history of; NOS, not otherwise specified; O/E, on examination.

descriptive data and because non-linear effects would be difficult to estimate over a 4-year period. A random effect for general practice was included because of the repeated observations on general practices over years. The log of person-time was included as offset. Relative rate reductions were estimated as one minus the adjusted relative rate for the linear effect of calendar year. In view of the size of the dataset, we present confidence intervals rather than significance tests. Results were presented using the 'ggplot2' and 'forest plot' packages²¹ in the R programme.²²

Research ethics

The research protocol for this study was submitted to and approved by the Medicines and Healthcare Products

Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC), protocol 16_020. All patients' electronic health records analysed for this study were fully anonymised.

Patient and public involvement

Neither patients nor public were involved in the development and design of this study, or the selection of outcome measures, or the conduct, analysis, and data dissemination of the study.

RESULTS

Overall antibiotic prescriptions

Analyses included 102 general practices that contributed data in each year from 2014 to 2017 (table 2). The registered population was 1.03 million in 2014 increasing to 1.07 in 2017. There were 539 219 AB prescriptions in 2014, declining to 459 476 in 2017. The AB prescribing rate declined from 608 per 1000 person-years in 2014 to 489 per 1000 person-years in 2017. The proportion of registered patients who were prescribed ABs in each year declined from just over 1 in 4 (25.3%) in 2014 to just over 1 in 5 (21.1%) in 2017. Figure 1 (left panel) shows changes in the proportion of patients prescribed ABs by year over the study period. A consistent year-on-year reduction was observed in each age-group from 0–4 years to ≥85 years. Marked AB prescribing variations were observed in relation to age, with the highest rates at the extremes of age.

A total of 195 750 broad-spectrum β -lactam AB prescriptions were made in 2014, declining to 153 423 in 2017. The proportion of all AB prescriptions that were broad-spectrum β -lactams decreased from 36.3% in 2014 to 33.4% in 2017 (table 2). Figure 1 (right panel) shows the change in proportion of patients prescribed broad-spectrum β -lactam ABs by age group. Although there was a year-on-year decrease in broad-spectrum β -lactam AB use in each age group, the absolute reduction appeared to be greater at older ages, where broad-spectrum β -lactam AB use was greatest.

Table 3 presents data for AB prescribing indications. Respiratory consultations accounted for the most frequent indication with 168 852 (31%) prescriptions in 2014 and 129 032 (28.1%) in 2017. Genitourinary infections and skin infections accounted for 9% and 7% of AB prescriptions, respectively, with little change over years. There were 77 431 (14%) AB prescriptions with no associated medical codes recorded in 2014 and 73 596 (16%) in 2017. There were 204 395 (38%) AB prescriptions with other and non-specific codes recorded in 2014 and 181 018 (39%) in 2017. Overall, more than half (54.1%) of the AB prescriptions were documented without specific clinical conditions recorded.

Table 4 shows the proportion of repeat prescriptions for different prescribing indications. In 2017, 78 166 (17%) AB prescriptions were recorded as repeat prescriptions. The proportion of repeat prescriptions was ≤2%



Table 2 Numbers of antibiotic (AB) prescriptions, and AB prescribing rates, by year. Figures are frequencies except where indicated.

	2014	2015	2016	2017
General practices	102	102	102	102
Patients	1 025 539	1 058 805	1 069 513	1 071 293
Female (%)	520 336 (50.7)	536 082 (50.6)	542 051 (50.7)	543 324 (50.7)
Age (mean, SD, years)	39.4 (23.4)	39.5 (23.4)	39.7 (23.5)	39.9 (23.5)
Person-time (person-years)	887 580	921 735	932 544	939 620
All AB prescriptions	539 219	494 185	482 917	459 476
All AB prescribing rate (per 1000 person-years)	608	536	518	489
Proportion of patients prescribed AB (%)	25.3	23.0	22.2	21.1
Mean number of AB prescriptions in patients receiving prescriptions	2.08	2.03	2.03	2.03
Broad-spectrum β -lactam AB prescriptions	195 750	174 353	167 056	153 423
Broad-spectrum β -lactam AB prescribing rate (per 1000 person-years)	221	189	179	163
Proportion of patients prescribed broad-spectrum β -lactam AB (%)	12.9	11.3	10.7	9.9
Mean number of broad-spectrum β -lactam AB prescriptions in patients prescribed	1.48	1.46	1.45	1.45

for respiratory, genitourinary, or eye conditions. For skin infections, 8% of AB prescriptions were recorded as repeat prescriptions. There were 10% of repeat prescriptions among AB prescribing associated with non-specific codes. Among 73 596 AB prescriptions in 2017 with no medical codes recorded, 56 216 (76%) were repeat prescriptions.

Informed by the apparent consistent annual declines in AB prescribing noted in table 2 and figure 1, figure 2 presents a Forest plot of annual relative reductions in AB prescribing adjusted for age, gender, and general practice. Estimates for all AB prescribing are shown in blue and for broad-spectrum β -lactam AB prescribing in red. The annual relative reduction in all AB prescribing was 6.9% (95% CI 6.6% to 7.1%). Estimates were generally similar for male and female subjects. For participants aged <55 years, the subgroup estimates were all greater

than the overall estimate, being greatest at age 45–54 years at 9.2% (8.4% to 9.9%) per year. For participants older than 55 years, estimates were consistently lower than the overall estimate being lowest at age 75–84 years and above at 4.3% (3.4% to 5.1%) per year. Considering subgroups of indications, rates of decline were greatest for respiratory indications (9.8%, 9.6% to 10.1%), and eye indications (11.0%, 9.9% to 12.2%). The rate of decline was smallest for AB prescriptions with no recorded indication (3.8%, 3.1% to 4.5%). The overall rate of decline was faster for broad-spectrum β -lactam ABs than all ABs at 9.3% (9.0% to 9.6%). Estimates were consistent for male and female subjects. The greatest relative decline was observed at 45–54 years (12.5%, 11.5% to 13.5%) and the lowest at 75–84 years (5.7%, 4.7% to 6.7%). The greatest decline was for skin condition indications (14.9%, 13.9% to 15.9%) and lowest for uncoded indications (5.5%, 4.5% to 6.4%).

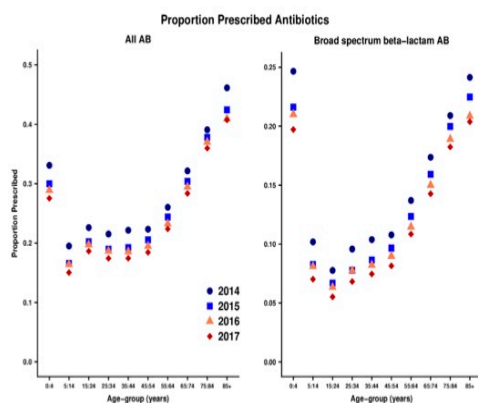


Figure 1 Proportion of patients prescribed antibiotics (ABs) in one year by age-group and calendar year.

Changes in different classes of antibiotics

Figure 3 presents changes over time in the use of different classes of ABs. The most frequently issued ABs were penicillins, accounting for 56% of AB prescriptions in men and 44% in women in 2017; macrolides, men 14%, women 12%; tetracyclines, men 14%, women 12%; sulfonamide and trimethoprim combination, men 6%, women 11%. Clindamycin, aminoglycosides, and other ABs accounted for <1% of AB prescriptions and are not shown. During the period of study, drugs for urinary tract infections (nitrofurantoin) increased as a proportion of all AB prescriptions, in men from 2.6% in 2014 to 4.2% in 2017, and in women from 8.8% in 2014 to 13.7% in 2017. Tetracycline use also increased between 2014 and 2017, in men from 12.8% to 14.5% and in women from

Table 3 Distribution of antibiotic (AB) prescriptions by broad groups of indications. Figures are frequencies except where indicated

	2014		2015		2016		2017		Total	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
AB prescriptions	539 219		494 185		482 917		459 476		1 975 797	
Respiratory conditions	168 852	31.3	146 025	29.5	140 263	29.0	129 032	28.1	584 172	29.6
Genitourinary conditions	47 009	8.7	44 544	9.0	42 453	8.8	42 401	9.2	176 407	8.9
Skin conditions	39 579	7.3	35 299	7.1	33 640	7.0	32 003	7.0	140 521	7.1
Eye conditions	1 953	0.4	1 622	0.3	1 586	0.3	1 426	0.3	6 587	0.3
Non-specific codes	204 395	37.9	191 565	38.8	189 386	39.2	181 018	39.4	766 364	38.8
No medical codes	77 431	14.4	75 130	15.2	75 589	15.7	73 596	16.0	301 746	15.3

10.1% to 11.6%. Most other categories appeared to show slight declines. Both penicillin and macrolides were mainly prescribed for treating respiratory conditions, whereas tetracyclines was frequently issued for skin conditions among young patients and respiratory conditions in later life. There was a decline in the use of sulfonamide/trimethoprim combinations for urinary conditions while a notable increase of nitrofurantoin use for these conditions was seen over the study years among all age groups, but more particularly in women.

Main findings

The rate of AB prescriptions and the proportion of patients receiving ABs have declined consistently over this 4-year period. Antibiotic use shows important variations by age and gender, being higher in very young and very old people and higher in women than men. However, these results show that a reduction in AB use is being achieved across all ages groups and in all subjects. The gender gap in relation to AB prescribing might be due to differences in medical care-seeking behaviour or specific conditions which disproportionately affect one gender.²³ Among prescriptions associated with coded indications, respiratory conditions were the most frequent indication for AB prescription and also showed the greatest rate of decline. Consistent with other recent reports,¹⁶ we find that a substantial proportion of AB prescriptions are not associated with specific coded clinical indications and of these, a major share is associated with repeat prescriptions. Antibiotic prescriptions that were not associated

with medical codes showed the slowest rate of decline, potentially further identifying this category of prescriptions as representing a suboptimal standard of clinical practice which might hamper the accurate estimation of drug indications. Therefore, enhancing the quality of clinical information recording is warranted in order to improve patient care, and the usefulness of records for research and health service management.

More than one-third of prescriptions were for β -lactam ABs and there was evidence of an important decline in AB prescribing in this category consistent with previous evidence.¹⁰ The relative reductions of broad-spectrum β -lactam prescriptions were greater than for overall AB use. Broad-spectrum β -lactam ABs may not necessarily offer more effective coverage of causal pathogens than their more specific counterparts. These results suggest that clinicians are gradually moving to more targeted narrow-spectrum substitutions when possible.

There is no universally accepted definition for 'broad-spectrum' ABs.^{10,19} This study analysed a separate category of β -lactam ABs that were broad-spectrum (as 'broad-spectrum β -lactam ABs') to illustrate the possible difference in prescribing trends between these broad-spectrum ABs and their counterparts. For most common and uncomplicated infections, narrower spectrum drugs are generally recommended as first-line agents in general practices.²⁴ Macrolides are generally recommended as substitutions for penicillin in cases of penicillin allergy, and for specific indications, including Legionella or the

Table 4 Proportion of antibiotic (AB) prescriptions that were either acute or repeat prescriptions in 2017. Figures are frequencies (percent of row total)

	Total AB prescriptions	Acute	Repeat
AB prescriptions	459 476	381 310 (83)	78 166 (17)
Respiratory conditions	129 032	127 474 (99)	1 558 (1)
Genitourinary conditions	42 401	41 740 (98)	661 (2)
Skin conditions	32 003	29 513 (92)	2 490 (8)
Eye conditions	1 426	1 399 (98)	27 (2)
Non-specific codes	181 018	163 804 (90)	17 214 (10)
No medical codes	73 596	17 380 (24)	56 216 (76)

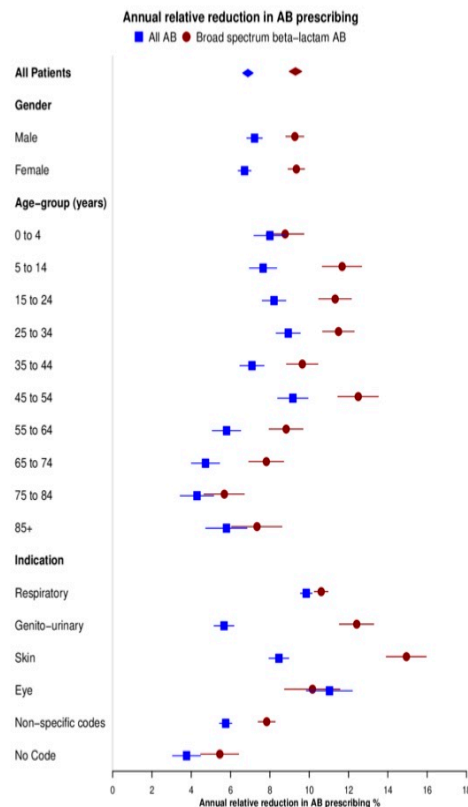


Figure 2 Forest plot showing annual relative reduction (95% CI) in antibiotic (AB) prescribing for all ABs and broad-spectrum β -lactam ABs between 2014 and 2017 for subgroups of age and gender and different prescribing indications. Estimates were adjusted for age, gender, and clustering by practice.

eradication of *Helicobacter pylori*. Nevertheless, macrolides were frequently prescribed in this and other studies.^{25 26} Clinical use of tetracyclines was low in children in recognition of the risk of deposition in growing bone and teeth,²⁷ but the overall use of tetracyclines was higher at other ages. The increase of nitrofurantoin use was mainly due to the change in the guideline recommendation from trimethoprim to nitrofurantoin as empiric treatment for urinary tract infection.²⁴

Strengths and limitations

The study included more than 100 general practices in England that participated consistently across the 4-year period of study. The CPRD includes general practices from throughout the UK. However, because the CPRD licence imposes limits on the size of dataset to be employed, we selected only CPRD general practices in England. During the period of the study, there was substantial attrition of the cohort of CPRD general practices as practices migrated from the Vision practice systems that were employed by practices contributing to the CPRD database. We considered that it was important to include the

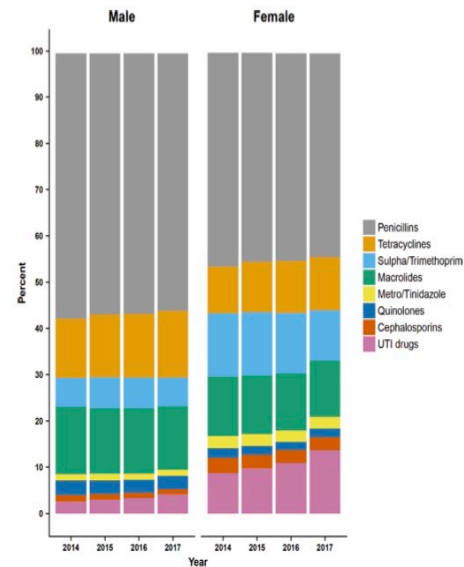


Figure 3 Bar chart showing changes from 2014 to 2017 in the proportion of antibiotic prescriptions for different antibiotic classes for male and female subjects. UTI, urinary tract infection.

same general practices in each year of study. However, we cannot be sure whether the AB prescribing of general practices that left the CPRD might differ from those that remained.

Previous studies have demonstrated the high quality and completeness of primary care electronic health records in CPRD.¹⁷ The data suggested that repeat AB prescriptions might account for a high proportion of uncoded prescriptions, but the prescription sequence field has not been well-validated to our knowledge. A concern for this study is the possible lack of recording of out-of-hours prescriptions, especially those from deputising services, walk-in centres, and emergency care settings.²⁸ We noted that codes for telephone consultations and home visits were frequent among AB prescriptions with non-specific coded indications, which suggests that some out-of-hours activity might have been captured. We also acknowledge that prescriptions from hospitals and specialist clinics are not included, but these are expected to make only a small contribution to community AB use. It appears unlikely that the large and consistent reductions in prescribing seen in this paper could be accounted for by movement of prescribing to other care settings.

The research analysed prescriptions issued and not prescriptions dispensed or consumed by patients. We could not determine whether prescribers used a delayed or deferred AB prescribing strategy. For these reasons, we believe that AB consumption may be slightly lower than we have reported. We acknowledge that there are variations in prescribing between practices.^{16 29 30} Our analytical method allowed us to estimate overall effects, and measures of precision, which took into account variation

between practices. Our results show some difference from an earlier study¹⁶ in the distribution of indications, but since different general practices, from different databases, were included in the two studies this may reflect variations in clinical practice.

Comparison with other studies

Previous analyses of primary care electronic health records have focused on AB prescribing for respiratory infections,^{31,32} recognising that these conditions represent the most frequent indications for AB prescription. There has been a long-term decline in respiratory consultation rates in England, which has contributed to reducing AB use for these conditions.³¹ Some authors suggest that respiratory consultations account for nearly two-thirds of AB use in primary care.³³ Our analyses are consistent with those of Dolk *et al.*¹⁶ who found that respiratory consultations account for fewer than half of AB prescriptions. However, a high proportion of prescriptions may be associated either with no medical codes or non-specific codes, making interpretation difficult. There were further methodological differences between the study of Dolk *et al.*¹⁶ and our own. The former study relied on the THIN database with a different number of general practices participating in different years, and used code lists that might have differed in some respects. Consequently, minor numerical differences are to be expected.

CONCLUSIONS

The present analyses add to recent reports by providing age- and gender-adjusted estimates of the rate of decline in AB use for all ABs and broad-spectrum β -lactam ABs, for different prescribing indications and different population subgroups defined by age and gender. The results show that the recent decline in AB use is broadly based and has occurred in all subgroups investigated. However, the decline in AB use has been at a faster rate for broad-spectrum β -lactam ABs than for all ABs; the decline is consistent by gender but tended to be lower over age 55 years; the slowest rate of decline is observed for AB prescriptions with no coded indications. The results emphasise the utility of electronic health records for providing individual-patient data for surveillance of trends in antimicrobial use and focusing future efforts at antimicrobial stewardship where these are most needed.

Contributors MCG and XS conceived the study. XS analysed and interpreted the data, MCG contributed additional analysis. XS wrote the draft of the manuscript and both authors revised and approved the final draft. XS is the guarantor.

Funding XS is supported by the China Scholarship Council. MCG was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London. This research is also supported by grants from the NIHR (HTA 13/88/10 and HS&DR 16/116/46).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Clinical Practice Research Datalink data were analysed under licence and are not available for sharing.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

REFERENCES

- Davies SC, Fowler T, Watson J, *et al.* Annual report of the chief medical officer: infection and the rise of antimicrobial resistance. *Lancet* 2013;381:1606–9.
- Review on Antimicrobial Resistance Chaired by Jim O'Neill. *Tackling drug-resistant infections globally: final report and recommendations*. London: Review on Antimicrobial Resistance, 2016.
- Ruiz J, Pons MJ, Gomes C. Transferable mechanisms of quinolone resistance. *Int J Antimicrob Agents* 2012;40:196–203.
- Gelband H, Pant S, Gandra S, *et al.* The state of the world's antibiotics, 2015. Washington D.C: Center for Disease Dynamics, Economics Policy, 2015 http://www.cddep.org/publications/state_worlds_antibiotics_2015#sthash.2fHwn4BD.dpbs
- Laxminarayan R, Duse A, Wattal C, *et al.* Antibiotic resistance—the need for global solutions. *Lancet Infect Dis* 2013;13:1057–98.
- Centers for Disease Control and Prevention. *Antibiotic resistance threats in the United States, 2013*. Atlanta, Georgia: Centres for Disease Control and Prevention, US Department of Health and Human Services, 2013.
- Costelloe C, Metcalfe C, Lovering A, *et al.* Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *BMJ* 2010;340:c2096.
- Goossens H, Ferech M, Vander Stichele R, *et al.* Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* 2005;365:579–87.
- World Health Organization. *Antimicrobial resistance. Global report on surveillance 2014*. Geneva, Switzerland: World Health Organization, 2014.
- Public Health England. *English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) Report 2017*. London: Public Health England, 2017.
- Public Health England. *English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) Report 2016*. London: Public Health England, 2016.
- Smith DRM, Dolk FCK, Pouwels KB, *et al.* Defining the appropriateness and inappropriateness of antibiotic prescribing in primary care. *J Antimicrob Chemother* 2018;73(suppl 2):ii11–18.
- van den Broek d'Obrenan J, Verheij TJM, Numans ME, *et al.* Antibiotic use in Dutch primary care: relation between diagnosis, consultation and treatment. *J Antimicrob Chemother* 2014;69:1701–7.
- Lusini G, Lapi F, Sara B, *et al.* Antibiotic prescribing in paediatric populations: a comparison between Viareggio, Italy and Funen, Denmark. *Eur J Public Health* 2009;19:434–8.
- NHS England. *Quality premium: 2016/17 Guidance for CCGs*. Leeds: NHS England, 2016.
- Dolk FCK, Pouwels KB, Smith DRM, *et al.* Antibiotics in primary care in England: which antibiotics are prescribed and for which conditions? *J Antimicrob Chemother* 2018;73(suppl 2):ii2–10.
- Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
- Herrett E, Shah AD, Boggon R, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;346:f2350.
- British Medical Association and Royal Pharmaceutical Society. *British National Formulary*. London: BMJ Group and Pharmaceutical Press, 2017.
- Rönnegård L, Shen X, Alam M. hglm: a package for fitting hierarchical generalized linear models. *R J* 2010;2:20–8.
- Wickham H. *ggplot2: elegant graphics for data analysis*. Heidelberg: Springer, 2016.
- Core Team R. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.
- Smith DRM, Dolk FCK, Smieszek T, *et al.* Understanding the gender gap in antibiotic prescribing: a cross-sectional analysis of English primary care. *BMJ Open* 2018;8:e020203.



24. Public Health England. *Management and treatment of common infections*. London: Public Health England, 2017.
25. Aabenhus R, Hansen MP, Saust LT, et al. Characterisation of antibiotic prescriptions for acute respiratory tract infections in Danish general practice: a retrospective registry based cohort study. *NPJ Prim Care Respir Med* 2017;27:37.
26. van den Broek d'Obrenan J, Verheij TJ, Numans ME, et al. Antibiotic use in Dutch primary care: relation between diagnosis, consultation and treatment. *J Antimicrob Chemother* 2014;69:1701–7.
27. National Institute for Health and Care Excellence. Tetracycline. London: NICE. 2018 <https://bnf.nice.org.uk/drug/tetracycline.html>
28. Williams SJ, Halls AV, Tonkin-Crine S, et al. General practitioner and nurse prescriber experiences of prescribing antibiotics for respiratory tract infections in UK primary care out-of-hours services (the UNITE study). *J Antimicrob Chemother* 2018;73:795–803.
29. Ashworth M, Charlton J, Ballard K, et al. Variations in antibiotic prescribing and consultation rates for acute respiratory infection in UK general practices 1995–2000. *Br J Gen Pract* 2005;55:603–8.
30. Pouwels KB, Dolk FCK, Smith DRM, et al. Explaining variation in antibiotic prescribing between general practices in the UK. *J Antimicrob Chemother* 2018;73(suppl_2):ii27–35.
31. Ashworth M, Latinovic R, Charlton J, et al. Why has antibiotic prescribing for respiratory illness declined in primary care? A longitudinal study using the General Practice Research Database. *J Public Health* 2004;26:268–74.
32. Gulliford M, Latinovic R, Charlton J, et al. Selective decrease in consultations and antibiotic prescribing for acute respiratory tract infections in UK primary care up to 2006. *J Public Health* 2009;31:512–20.
33. National Institute for Health and Care Excellence. *Prescribing of antibiotics for self-limiting respiratory tract infections in adults and children in primary care*. London: National Institute for Health and Clinical Excellence, 2008.

Pneumonia incidence trends in UK primary care from 2002 to 2017: population-based cohort study

Original Paper

Cite this article: Sun X, Douiri A, Gulliford M (2019). Pneumonia incidence trends in UK primary care from 2002 to 2017: population-based cohort study. *Epidemiology and Infection* **147**, e263, 1–7. <https://doi.org/10.1017/S0950268819001559>

Received: 13 February 2019




Revised: 30 July 2019

Accepted: 12 August 2019

Key words:

Antibiotics; epidemiology; pneumonia; primary care; respiratory tract infection

Author for correspondence: Xiaohui Sun, E-mail: xiaohui.sun@kcl.ac.uk

Xiaohui Sun¹ , Abdel Douiri^{1,2}  and Martin Gulliford^{1,2} 

¹King's College London, School of Population Health and Environmental Sciences, London, UK and ²National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' National Health Service Foundation Trust, London, UK

Abstract

Increasing hospital admissions for pneumonia have been reported recently but it is not known whether pneumonia incidence rates have increased in the community. To determine whether incidence rates of pneumonia increased in primary care in the United Kingdom from 2002 to 2017, an open cohort study was conducted using electronic health records from the UK Clinical Practice Research Datalink. Clinically diagnosed pneumonia, influenza pneumonia, pleural infection and clinically suspected pneumonia, defined as chest infection treated with antibiotics, were evaluated. Age-standardised and age-specific rates were estimated. Joinpoint regression models were fitted and annual percentage changes (APC) were estimated. There were 70.7 million person-years of follow-up with 120 662 episodes of clinically diagnosed pneumonia, 1 831 005 of clinically suspected pneumonia, 23 814 episodes of influenza pneumonia and 2644 pleural infections over 16 years. The incidence of clinically diagnosed pneumonia increased from 1.50 per 1000 person-years in 2002 to 2.22 per 1000 in 2017. From 2010 to 2017, the APC in age-standardised incidence was 5.1% (95% confidence interval 3.4–6.9) compared with 0.3% (–0.6 to 1.2%) before 2010. Clinically suspected pneumonia incidence rates increased from 2002 to 2008 with an APC 3.8% (0.8–6.9) but decreased with an APC –4.9% (–6.7 to –3.1) from 2009 to 2017. Influenza pneumonia increased in the epidemic year of 2009. There was no overall trend in pleural infection. The results show that clinically diagnosed pneumonia has increased in primary care but there was a contemporaneous decline in recording of clinically suspected pneumonia or ‘chest infection’. Changes in disease labelling practice might partly account for these trends.

• What is the key question?

Recent evidence suggests an increasing trend in pneumonia hospitalizations. This study analysed primary care electronic health records for more than 70 million patient years of follow-up to evaluate whether pneumonia incidence had increased in community settings.

• What is the bottom line?

Clinically-diagnosed pneumonia increased from 2002 to 2017, with an acceleration in trend after 2011. There was a simultaneous decrease in the more frequent diagnosis of clinically-suspected pneumonia characterized as antibiotic-treated chest infection. This suggests that changes in diagnostic labelling may at least in part account for the apparent increase in clinically-diagnosed pneumonia.

• Why read on?

The large study, including practices from throughout the UK, sheds new light on previous reports of increasing pneumonia hospitalisations. The study shows that in adults there have been divergent trends in different respiratory infection diagnoses with increasing clinically-diagnosed pneumonia and reducing clinically-suspected pneumonia. This is in contrast to consistently decreasing lower respiratory infections in children.

Introduction

With the advent of antibiotics, common but severe infections such as pneumonia could be effectively cured with access to effective antimicrobial treatment [1] but community acquired pneumonia (CAP) remains a major public health priority worldwide, disproportionately affecting younger and older populations [2–4]. As an ambulatory care sensitive condition, [5] pneumonia may be managed in primary care or may result in hospital admission, depending on assessment of severity [6]. Recently, increasing hospital admissions for pneumonia have been reported in several studies [2, 7, 8]. In the USA, Fry *et al.* [2] reported increasing pneumonia hospitalisation rates in people aged 65 years and older and suggested that increasing comorbidity might be a contributing factor. In England, analysis of hospital episode statistics

© The Author(s) 2019. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

CAMBRIDGE
UNIVERSITY PRESS

showed increasing hospital admissions for pneumonia from 1997 to 2005 [7]. The increase did not appear to be fully explained by demographic change, nor by an increase in co-existing conditions [7]. In Oxfordshire, hospital admissions for CAP increased from 1998 to 2014, with a more rapid rate of increase after 2008 [8]. Trends were similar in all age groups [8]. Other studies also suggest that the rate of emergency hospital admissions for pneumonia is increasing [9, 10].

While accumulating evidence points to a growing burden of pneumonia on hospital services, this may not necessarily mean that the incidence of pneumonia disease is increasing in the general population. Patients with pneumonia treated in hospital settings may differ from those managed in community. The decision to admit to hospital may depend on a patient's general health condition and social circumstances, as well as severity of illness [11]. Pneumonia may complicate pre-existing illnesses that make individuals vulnerable to infectious pathogens [12, 13]. Pneumonia itself may contribute to deterioration in underlying or pre-existing medical conditions including heart failure [14] or coronary syndromes [15] for which hospitalisation may be indicated. Investigation of the epidemiology of respiratory syndromes in the community is therefore indicated.

In primary care, management of respiratory conditions may often be 'symptom oriented'. Definitive diagnosis through confirmatory tests may be less readily available in comparison with secondary care. Family physicians in the UK use the Snomed CT [16] and Read Code classifications [17], which enable coding of comprehensive and detailed patient information including occupation, social circumstances, clinical symptoms and signs, clinical tests and use of medical services [16]. This is in contrast to the disease categorisations offered by the International Classification of Diseases (ICD) [18] used for coding of hospital episode statistics. Primary care classifications provide detailed and granular coding systems including a wider range of information to enable patient management in the community. In primary care, CAP cases may often be identified based on clinical features rather than from clinical investigations including radiology findings and bacteriological tests [11]. Less specific labels including 'chest infection' may be applied to clinically suspected pneumonia consultations. Clinical guidelines in the UK emphasise that the category of 'chest infection' may contain two general clinical scenarios: acute bronchitis and CAP, with antibiotics only being indicated for the latter [19]. Conversely, there may be no major differences between management recommendations for CAP and clinically suspected pneumonia labelled as 'chest infection' in terms of essential elements of treatment, severity assessment and referral principles in adult population [6, 19]. From a disease management perspective, adult CAP and antibiotic-treated chest infection cases might be labelled interchangeably during consultations in primary care.

To provide complementary information to hospital-based studies, this study aims to evaluate trends in the incidence of conditions managed as pneumonia in the community. In this study, CAP was included as 'clinically diagnosed pneumonia' whereas antibiotic-treated chest infection as 'clinically suspected pneumonia'. Influenza pneumonia was included to assess whether secondary pneumonia complicating influenza has exerted a significant impact on pneumonia burden in the community [20, 21]. Pleural infection was also analysed because it shares a similar aetiology to pneumonia and may be associated with pneumonia severity, with recent studies documenting increasing pleural infection trends especially among children [22].

Methods

Data source and study design

A population-based cohort study was conducted in the UK Clinical Practice Research Datalink (CPRD). The CPRD is the world's largest database of primary care electronic health records, covering approximately 7% of the UK family practices [23]. More than 98% of the UK population were registered with a family practice [24]. The CPRD population is considered to be representative of the UK population; the database includes comprehensive data for drug prescriptions and diagnoses recorded in primary care, which have been shown to be valid in many studies [23, 24]. The CPRD comprises an open cohort of UK family practices and their registered patients. The present analysis included all eligible family practices contributing to CPRD, and data for all registered patients aged up to 100 years old, during 16 calendar years from the beginning of 2002 to the end of 2017.

Cases

We included recorded diagnoses of pneumonia, antibiotic-treated chest infection, influenza pneumonia and pleural infection, including bacterial pleurisy and empyema. Read codes were reviewed and selected independently by two researchers with clinical and epidemiological backgrounds. Pneumonia was defined using Read codes associated with 'pneumonia' terms after excluding tuberculosis (TB), fungal and parasite pneumonia. Influenza pneumonia was evaluated as a separate group. The remaining pneumonia codes were grouped into a single category of 'bacterial pneumonia' after excluding non-infectious pneumonia codes e.g. bronchiolitis obliterans organising pneumonia [25]. This was consistent with the category of CAP being mainly used to refer to uncomplicated bacterial pneumonia in the general population. CAP was evaluated as 'clinically diagnosed pneumonia' in this study. Antibiotic-treated chest infection cases were identified when patients were recorded with 'chest infections' and received antibiotic prescription on the same day. 'Antibiotics' included antibacterial agents from chapter 5.1 of the British National Formulary without anti-viral, anti-TB, anti-leprosy and anti-fungal drugs [26]. Antibiotic-treated chest infection was then analysed as clinically suspected pneumonia. Empyema and bacterial pleurisy were analysed as pleural infection. Primary diagnoses were not differentiated when multiple diagnoses were recorded in a single consultation. Codes for the same condition in the same patient during a 90-day time-window were considered to represent a single episode.

Statistical analysis

Person-time at risk was estimated for the CPRD registered population by year from 2002 to 2017 as a denominator. For each patient, we included time from the latest of the patient registration date, or the date the family practice began contributing data to CPRD, to the earliest of the patient end-of-registration, death date or the date the practice left CPRD. Incident events were considered as those recorded more than 1 year after the patient start-date to eliminate prevalent cases from any possible duplication of records during patient registration. Age-specific incidence rates were calculated using the age-groups 0–4 years, 5–14 years, then 10-year age-groups, up to 85 years and older. Data for participants aged more than 100 years were omitted as these are few in number and may be subject to data recording errors. Age- and sex-standardised incidence rates (ASRs) were calculated using the European standard population

Table 1. Number of incidence events of pneumonia and related conditions

Year	Family practices	Person years	Clinically diagnosed pneumonia		Clinically suspected pneumonia		Influenza pneumonia		Pleural infection	
			Freq.	Rate ^a	Freq.	Rate ^a	Freq.	Rate ^a	Freq.	Rate ^a
2002	550	4 191 630	6291	1.50	99 256	23.68	956	0.23	136	0.03
2003	586	4 458 959	7189	1.61	116 781	26.19	1081	0.24	157	0.04
2004	612	4 767 931	7219	1.51	117 265	24.59	1264	0.27	157	0.03
2005	620	4 898 961	8184	1.67	139 785	28.53	1473	0.30	186	0.04
2006	626	4 956 288	7812	1.58	134 030	27.04	1454	0.29	192	0.04
2007	631	5 016 169	7798	1.55	151 411	30.18	1468	0.29	200	0.04
2008	627	5 025 191	8043	1.60	152 922	30.43	1500	0.30	207	0.04
2009	621	5 026 729	7977	1.59	133 848	26.63	6103	1.21	215	0.04
2010	613	4 967 771	8145	1.64	136 905	27.56	2110	0.42	210	0.04
2011	596	4 862 957	8141	1.67	127 963	26.31	1574	0.32	177	0.04
2012	580	4 805 309	8529	1.77	133 994	27.88	1309	0.27	196	0.04
2013	564	4 595 318	8376	1.82	112 730	24.53	1025	0.22	177	0.04
2014	530	4 201 387	7802	1.86	96 219	22.90	859	0.20	134	0.03
2015	462	3 605 006	7353	2.04	76 108	21.11	672	0.19	107	0.03
2016	371	2 861 175	6346	2.22	57 126	19.97	536	0.19	102	0.04
2017	314	2 455 307	5457	2.22	44 662	18.19	430	0.18	91	0.04

Figures are frequencies except where indicated.
^aRate per 1000 person years.

(2013 revision). To evaluate whether there had been any recent changes in trend, changes over calendar year were modelled by joinpoint regression analysis [27] using Joinpoint Trend Analysis software from the NIH National Cancer Institute [28]. The joinpoint method starts with a linear model and tests whether addition of joinpoints improves goodness of fit using Monte-Carlo permutation tests [27]. Annual percentage changes (APCs) were estimated to quantify the direction and slope of the trend in given period of time and average annual percent change (AAPC) was adopted to measure average rate changes across the whole study period.

Sensitivity analyses

To evaluate whether changes in the family practice population influenced conclusions, we repeated analyses including only the 218 family practices that contributed data in each year of study from 2002 to 2017.

Research ethics

The research protocol for this study was submitted to and approved by the Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC), Protocol 16_020. All patient medical records were anonymised before data received by researchers.

Results

There were 550 family practices contributing to CPRD in 2002, increasing to 631 in 2007, before declining to 314 in 2017

(Table 1). There was a total of 4.2 million person-years of follow-up of registered patients aged up to 100 years in 2002, increasing to 5.03 million in 2008–2009, before declining to 2.5 million in 2017.

Clinically diagnosed pneumonia

The number of episodes of clinically diagnosed pneumonia was between 5000 and 10 000 in each year of study (Table 1). Six codes accounted for 81% of all pneumonia episodes: 'pneumonia due to unspecified organism' (38%); 'bronchopneumonia due to unspecified organism' (14%); 'history of pneumonia' (12%); 'community acquired pneumonia' (8%); 'lobar (pneumococcal) pneumonia' (7%) and 'lobar pneumonia due to unspecified organism' (3%). The crude incidence rate of clinically diagnosed pneumonia increased from 1.50 (1.46–1.54) per 1000 patient years in 2002 to 1.64 (1.60–1.68) per 1000 in 2010, the rate then increased more rapidly to 2.22 (2.16–2.28) per 1000 in 2017. Figure 1 (top left panel) shows changes in the age-standardised rates of pneumonia for men and women separately with fitted lines from joinpoint regression. Trends were similar in men and women but clinically diagnosed pneumonia was more frequent in men. Table 2 presents estimates from the joinpoint regression model. The APC in age-standardised rate of clinically diagnosed pneumonia was 0.3% (95% confidence interval –0.6 to 1.2) per year from 2002 to 2010 but from 2010 to 2017 the APC was 5.1% (3.4–6.9) per year. The AAPC over the entire period was 2.4% (3.4–6.2) per year. Estimation of age-specific rates of pneumonia (Fig. 2) shows that clinically diagnosed pneumonia was reducing throughout the period in children aged under 15 years, while recorded

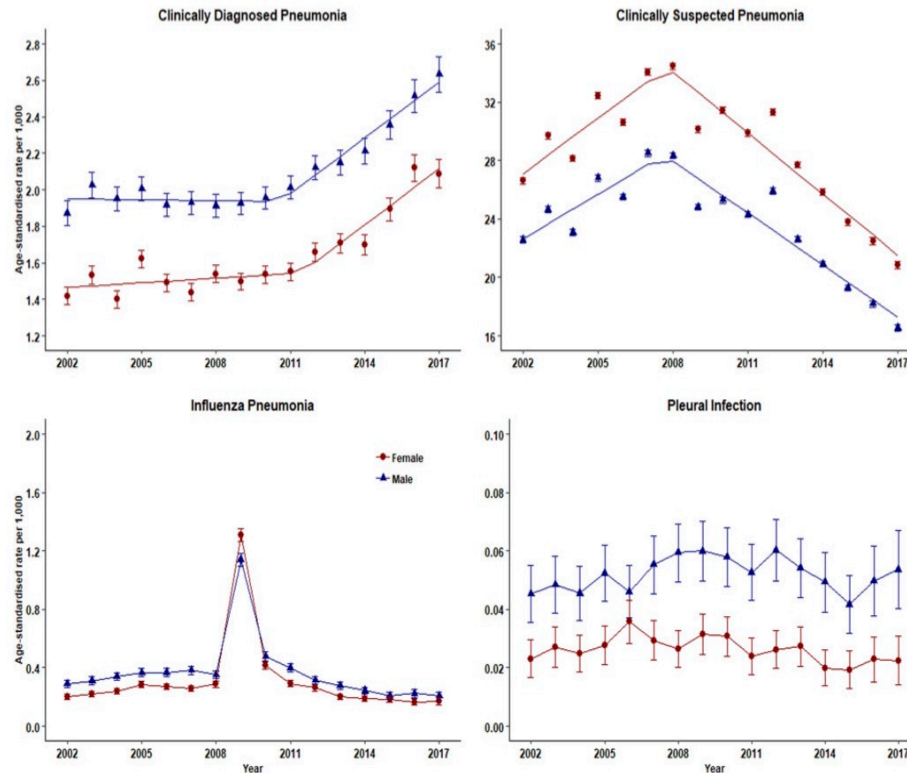


Fig. 1. Trends in pneumonia and related conditions for both men (blue) and women (red) 2002–2017. Rates are per 1000 person-years.

Table 2. Joinpoint regression estimates for APC

Condition	Measure	Year of joinpoint	APC (%) before joinpoint (95% CI)	APC (%) after joinpoint (95% CI)	Average APC (%) 2002 to 2017 (95% CI)
Clinically diagnosed pneumonia	Crude-rate	2010	0.4 (−0.7 to 1.5)	4.8 (3.4–6.2)	2.4 (3.4–6.2)
	ASR	2011	0.3 (−0.6 to 1.2)	5.1 (3.4–6.9)	2.2 (1.4–3.0)
Clinically suspected pneumonia	Crude-rate	2008	3.9 (0.9–7.1)	−4.7 (−6.6 to −2.9)	−1.4 (−2.8 to 0.1)
	ASR	2008	3.8 (0.8–6.9)	−4.9 (−6.7 to −3.1)	−1.5 (−2.9 to −0.1)

clinically diagnosed pneumonia increased in adults, especially at older ages. In patients aged 15–54 years, rates of clinically diagnosed pneumonia were slightly higher in women than men but, but over the age of 55 years clinically diagnosed pneumonia was more frequent in men, especially at the oldest ages.

Clinically suspected pneumonia

The annual number of cases of clinically suspected pneumonia ranged between 44 662 and 152 992. Two codes accounted for more than 99% of clinically suspected pneumonia: ‘chest infection not otherwise specified’ (61%) and ‘chest infection’ (39%). The crude rate of clinically suspected pneumonia was more than 10 times higher than for clinically diagnosed pneumonia, increasing from 23.7 in 2002 to 30.4 per 1000 in 2008 before declining to

18.2 per 1000 in 2017 (Table 1). Figure 1 (top right panel) shows that trends age-standardised rates of clinically suspected pneumonia were similar in men and women, but absolute rates were greater in women than in men. Joinpoint regression indicated a change in trend in 2008. The APC from 2002 to 2008 was 3.8% (0.8–6.9) per year compared with −4.9% (−6.7 to −3.1) per year after 2008. Changes in age-specific rates were generally consistent but clinically suspected pneumonia was more frequent in women from 15 to 74 years but more frequent in males during childhood and over the age of 75 years. The overall trend for all chest infection diagnoses was similar to clinically suspected pneumonia with the same turning point of 2008. While the proportions of all chest infections that were clinically suspected pneumonia increased steadily from 66% in 2002 to 88% in 2017 at an average APC of 1.6% (1.4–1.8) per year, suggesting

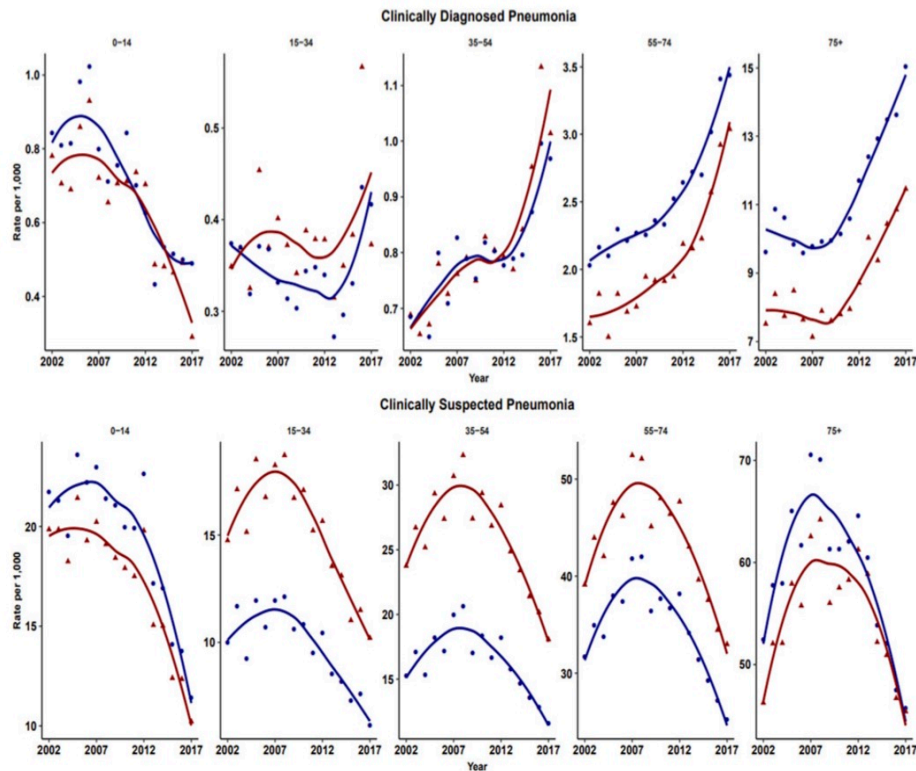


Fig. 2. Age-specific rates for clinically diagnosed pneumonia and clinically suspected pneumonia for males (blue) and females (red). Rates are per 1000 person-years.

that 'chest infection' not treated with antibiotics declined more rapidly than treated chest infection.

Influenza pneumonia and pleural infection

Table 1 and Figure 1 present data for influenza pneumonia and pleural infection (including bacterial pleurisy and empyema). Rates of influenza pneumonia showed a peak in 2009 but remained low in other years. Rates of pleural infection were low and showed no consistent trend over time. Pleural infection was more frequent in men than women but there was no gender difference for influenza pneumonia.

Sensitivity analysis

To evaluate whether attrition of family practices from CPRD, accounted for changes in coding, analyses were repeated using only data from 218 family practices that contributed data in every year from 2002 to 2017. In these 218 practices, the crude rate of clinically diagnosed pneumonia increased from 1.38 per 1000 in 2002 to 1.56 per 1000 in 2010 and then increased to 2.24 per 1000 in 2017 with an APC being 3.1% (1.9–4.2). For clinically suspected pneumonia, the crude rate increased from 20.8 per 1000 in 2002 to 29.7 per 1000 in 2008 at an APC of 4.0% (0.9–7.2) before declining to 18.0 per 1000 in 2017 at an average percentage of 5.1% (–6.6% to –3.4).

Discussion

Main findings

There was an increasing trend in clinically diagnosed pneumonia from 2002 onwards and this accelerated after 2011. This was unlikely to be due population ageing because similar trends were observed for age-standardised and crude incidence rates. Analysis of clinically suspected pneumonia showed that this syndrome was much more frequently recorded than clinically diagnosed pneumonia and incidence rates increased from 2002 to 2008 but decreased rapidly thereafter. Clinicians necessarily work with diagnostic disease classifications but pulmonary infections may represent graduated phenomena with varying degrees of bronchial inflammation and alveolar consolidation. This may contribute to diagnostic uncertainties and perhaps inconsistent selection of diagnostic terms. Given that 'chest infection' may not be a confident diagnosis, together with the volume of antibiotic-treated chest infection rates being considerably higher than that of diagnosed pneumonia, a small change in disease coding practice could lead to a shift from 'chest infection' recording to 'pneumonia' recording. Joinpoint regression analysis suggested that the decline in 'chest infection' recording began in 2008 slightly before the increase in 'pneumonia' recording from 2011. However, conditions managed as pneumonia were considered as relatively stable as none of the AAPCs have shown to be statistically significant from middle age and above. This would suggest that there was code drifting during clinical consultations when

pneumonic infectious symptoms were presented among adult population with elder patient being more likely being diagnosed with pneumonia.

According to UK guidelines, adult 'chest infection' should be managed as CAP when pneumonia is suspected [20]. However, Petersen *et al.* [29] regarded pneumonia as a potential complication of chest infection in their electronic health records based study. If antibiotics are prescribed less frequently for chest infection it is possible that pneumonia might increase [30]. One study suggested that pneumonia might be more frequent at family practices that prescribe fewer antibiotics for respiratory infections [30]. Thus although it appears likely that changes in coding of respiratory infections account for observed trends, the present data do not exclude the possibility that changing management of 'chest infections' is leading to an increase in pneumonia incidence. Therefore, understanding the underlying reason for adopting certain disease labels such as 'chest infection' rather than confident disease diagnoses during routine healthcare in the community would contribute to understanding the challenges of diagnostic uncertainty in primary care settings.

In children, records of both clinically diagnosed pneumonia and clinically suspected pneumonia decreased. This could be explained by the introduction of *Haemophilus Influenzae* Type B vaccine in 1992 and pneumococcal conjugate 7 vaccine in 2006 into the UK childhood immunisation scheme [31]. In adults, rates of clinically diagnosed pneumonia increased while clinically suspected pneumonia decreased. Trends were generally similar in males and females but women between the ages of 15 and 74 years were more likely to be recorded with antibiotic-treated chest infection than men, but this distinction was not apparent for clinically diagnosed pneumonia. It is unclear whether this represents a disease classification preference or whether more severe cases tended to be found in men. Influenza pneumonia showed an increase in the epidemic year of 2009 [21] but overall low incidence of influenza pneumonia might result from the influenza vaccination schemes among both young and elder populations [31]. Pleural infection was analysed as a surrogate of severe pneumonia because recent evidence suggests an increase in incidence rate trends among children [22]. Our results showed that the incidence trends for pleural infection remained stable during the past 16 years.

Comparison with previous studies

In contrast to previous studies on pneumonia burden using hospital admission data [7, 8], this study using electronic health records data with general practitioners being data recorders. Since pneumonia patients referred to secondary care may not necessarily be representative of all CAP, using primary care consultation data contributes to understanding pneumonia patients presenting and managed in primary care settings.

In National Institute for Health and Care Excellence (NICE) guidelines for both CAP and chest infection, together with British Thoracic Society recommendations for CAP management in adult patients, CRB-65 score (confusion, raised respiratory rate, low blood pressure and age 65 and above) is recommended to guide risk assessment and place of treatment [6, 19, 32]. This score identifies older age (≥ 65 years old) as an independent risk score since being 65 years and above will automatically classify patients into an intermediate-risk group. This implies that

more conservative management strategies have been applied to older pneumonia patients. This partially explained previous study findings that pneumonia patients in the community leading to hospitalisation were increasing in recent years among elder populations and there is no evidence that less severe patients were admitted to secondary care.

Strengths and limitations

This population-based study analysed healthcare records to outline the range of conditions that were managed as pneumonia in the community. The 16-year timeframe provided sufficient data to estimate disease trends over a substantial study period. The study included all eligible practices and patients contributing health care information to CPRD more than 1 year. There were 320 practices participating the database by the end of 2017. Previous studies have shown the completeness and data of high quality in CPRD. The large sample size of research cohort was sufficient for depicting the trends for clinically diagnosed pneumonia, clinically suspected pneumonia, even low incidence conditions such as influenza pneumonia and pleural infections.

CPRD has not released free text information since 2013, which would enable us to examine diagnostic information documented in free text records rather than coded. But, we consider that clinicians would record relevant conditions such as pneumonia or chest infection especially when clinical discretion leads to antibiotic treatment. Also, health care information in CPRD made disease severity assessment unfeasible, therefore, we could not confidently determine whether case severity influenced coding practices or place of treatment. Also, we did not capture data from out-of-hour services, walk-in centre consultations and emergency care.

This study was derived from a universal health care coverage system where most common conditions are managed in primary care settings, implications generated from this study would mainly apply to similar systems but not where health care insurance plays an essential role or referral thresholds from primary care to secondary care vary extensively compared with that in the UK.

Conclusion

Clinically diagnosed pneumonia is increasing over time in the UK. This trend could not be fully explained by ageing population, changes in coding practice or alternative diagnosis. Age-specific trends were divergent with decreasing trends in children but increasing in older adults. Respiratory conditions managed as pneumonia in family practice were decreasing slightly over time, which was more likely due to more conservative antibiotic prescribing strategies. Research to reduce diagnostic uncertainty would contribute to improving antibiotic stewardship in the community.

Acknowledgement. AD and MG acknowledge support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London. MG's research is also supported by grants from the NIHR (HTA 13/88/10 and HS&DR 16/116/46).

Conflict of interest. None.

Funding. XS is supported by King's-China Scholarship Council.

References

- Podolsky SH (2005) The changing fate of pneumonia as a public health concern in 20th-century America and beyond. *American Journal of Public Health* **95**, 2144–2154.
- Fry AM *et al.* (2005) Trends in hospitalizations for pneumonia among persons aged 65 years or older in the United States, 1988–2002. *JAMA* **294**, 2712–2719.
- Jain S *et al.* (2015) Community-acquired pneumonia requiring hospitalization among US children. *New England Journal of Medicine* **372**, 835–845.
- Troeger C *et al.* (2018) Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Infectious Diseases* **11**, 1191–1120.
- Frick J *et al.* (2017) Suitability of current definitions of ambulatory care sensitive conditions for research in emergency department patients: a secondary health data analysis. *BMJ Open* **7**, e016109.
- National Institute for Health and Care Excellence (2014) *Pneumonia in Adults: Diagnosis and Management*. London: NICE.
- Trotter CL *et al.* (2008) Increasing hospital admissions for pneumonia, England. *Emerging Infectious Diseases* **14**, 727.
- Quan TP *et al.* (2016) Increasing burden of community-acquired pneumonia leading to hospitalisation, 1998–2014. *Thorax* **71**, 535–542.
- Blunt I (2013) *Focus on Preventable Admissions*. London: Nuffield Trust.
- Bardsley M *et al.* (2013) Is secondary preventive care improving? Observational study of 10-year trends in emergency admissions for conditions amenable to ambulatory care. *BMJ Open* **3**, e002007.
- Lim WS *et al.* (2009) British thoracic society guidelines for the management of community acquired pneumonia in adults: update 2009. *Thorax* **64**(Suppl. 3), iii1–iii55.
- Lai S-W, Lin C-L and Liao K-F (2017) Risk of pneumonia among patients with splenectomy: a retrospective population-based cohort study. *Annals of Saudi Medicine* **37**, 351–356.
- Lai S-W, Lin C-L and Liao K-F (2017) Risk of contracting pneumonia among patients with predialysis chronic kidney disease: a population-based cohort study in Taiwan. *Biomedicine* **7**.
- Eurich DT *et al.* (2017) Risk of heart failure after community acquired pneumonia: prospective controlled study with 10 years of follow-up. *BMJ* **356**, j413.
- Corrales-Medina VF, Madjid M and Musher DM (2010) Role of acute infection in triggering acute coronary syndromes. *The Lancet Infectious Diseases* **10**, 83–92.
- NHS Digital (2018) *Snomed CT*. Leeds: NHS Digital. Available at <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct> (Accessed 21 May 2019).
- NHS Information Authority (2004) *Clinical Terms (Read Codes). Summarised Product Description*. Leeds: NHS Information Authority.
- World Health Organization (2018) *International Statistical Classification of Diseases and Related Health Problems 10th Revision 2010*. Geneva: World Health Organization. Available at <https://icd.who.int/browse10/2010/en> (Accessed 21 May 2019).
- National Institute for Health and Care Excellence (2015) *Chest Infections – Adult*. London: National Institute for Health and Care Excellence.
- Metersky ML *et al.* (2012) Epidemiology, microbiology, and treatment considerations for bacterial pneumonia complicating influenza. *International Journal of Infectious Diseases* **16**, E321–E331.
- Chan M (2009) *World now at the Start of 2009 Influenza Pandemic*. Geneva: World Health Organisation. Available at https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/ (Accessed 21 May 2019).
- Mahon C *et al.* (2016) Incidence, aetiology and outcome of pleural empyema and parapneumonic effusion from 1998 to 2012 in a population of New Zealand children. *Journal of Paediatrics and Child Health* **52**, 662–668.
- Williams T *et al.* (2012) Recent advances in the utility and use of the general practice research database as an example of a UK primary care data resource. *Therapeutic Advances in Drug Safety* **3**, 89–99.
- Herrett E *et al.* (2015) Data resource profile: clinical practice research datalink (CPRD). *International Journal of Epidemiology* **44**, 827–836.
- Epler GR (2001) Bronchiolitis obliterans organizing pneumonia. *Archives of Internal Medicine* **161**, 158–164.
- Joint Formulary Committee (2017) *British National Formulary 74: September 2017*. London, UK: Pharmaceutical Press.
- Kim HJ *et al.* (2000) Permutation tests for jointpoint regression with applications to cancer rates. *Statistics in Medicine* **19**, 335–351.
- National Cancer Institute (2018) *Joinpoint Regression Program [Internet]*. Bethesda: National Cancer Institute. Available at <https://surveillance.cancer.gov/joinpoint/> (Accessed 21 May 2019).
- Petersen I *et al.* (2007) Protective effect of antibiotics against serious complications of common respiratory tract infections: retrospective cohort study with the UK general practice research database. *BMJ (Clinical Research Ed)* **335**, 982.
- Gulliford MC *et al.* (2016) Safety of reduced antibiotic prescribing for self limiting respiratory tract infections in primary care: cohort study using electronic health records. *BMJ* **354**, i3410.
- Public Health England (2013) *Historical Vaccine Development and introduction of Routine Vaccine Programmes in the UK*. London: Public Health England.
- British Thoracic Society (2015) *Annotated BTS Guideline for the Management of CAP in Adults. Summary of Recommendations*. London: British Thoracic Society.

