AI and the Sacred Disease – The Opportunities of Electronic Patient Records and Natural Language Processing to Advance Epilepsy Care and Beyond

Shek, Anthony

Awarding institution:
King's College London

AI and the Sacred Disease – The Opportunities of Electronic Patient Records and Natural Language Processing to Advance Epilepsy Care and Beyond

Shek, Anthony

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Clinical Neuroscience

Department of Basic and Clinical Neuroscience
Institute of Psychiatry, Psychology and Neuroscience (IoPPN)
KING'S COLLEGE LONDON
UNITED KINGDOM

March 2022

Abstract

Patients with epilepsy frequently visit healthcare providers, making routinely collected data ideal sources of information about their health status. Each patient-healthcare provider interaction is recorded in the Electronic Patient Record (EPR). These data range from documentation of diagnoses and symptoms to procedures, prescriptions, and tests. This granularity of data makes EPRs ideal sources of information to identify disease patterns and provide evidence for the optimisation of health management. However, beyond direct patient care, the use of EPR data for secondary purposes such as research or service development has so far been limited. This is because the majority of clinically valuable information is contained within unstructured or free-text documentation. In this form, the lack of format requirements permits the use of homonyms, spelling mistakes, abbreviations, and use of legacy epilepsy classification terminology. For this information to be extracted and grouped for downstream machine-interpretable analytics, these details must be standardised into distinct areas of meaning.

This thesis aims to provide a fresh insight into correlations between patient care pathways and specific outcomes of people with epilepsy, at a scale not previously possible with traditional study designs or manually created data sources. This will be done by centralising, capturing, and aggregating large amounts of structured and unstructured clinical narratives from heterogeneous textual data sources, i.e. different types of medical records and administrative records from a secondary care setting – King's College Hospital NHS foundation trust (KCH) – using an information retrieval and extraction platform called CogStack.

We begin by exploring first suspected seizure patients and how they were managed at a time when there was no formal first fit care pathway. We found that approximately half of cases were not in keeping with NICE guidelines. We used rules-based natural language processing (NLP) techniques to uncover that the most commonly reached diagnosis after a first suspected seizure event were epilepsy, cardiovascular disorders, and a single unknown episode.

This thesis then proceeds to describe the development of MedCAT, an NLP tool that uses a novel self-supervised machine learning algorithm approach for identifying clinical information and linking them to clinical concepts from a standardised terminology. Real-world validation demonstrates accurate SNOMED-CT concept extraction from different EPR vendor systems at 3 large London hospitals with self-supervised training over 8.8 billion words from >17 million clinical records and further supervised training fine-tuning step with ~6,000 clinician or domain expert-annotated examples. We discuss the transferability of models and reveal retained performances, across different hospitals and datasets, to allow the re-use and site-specific finetuning of models.

We then demonstrate the application of the CogStack-MedCAT pipeline to gather a large retrospective epilepsy cohort of 4,011 patients. These patients were then subgrouped based on demographics, epilepsy type, and comorbidity to provide insights into healthcare service utilisation, anti-seizure medication treatment patterns, and health outcomes. At KCH, the majority of patients were classified under an unknown epilepsy type (43.3%) followed by focal (37.3%), generalised (15.5%), and lastly combined generalised and focal epilepsy (3.9%). In terms of anti-seizure medication precipitation patterns, we found that levetiracetam was the most popular monotherapy (17.3%) and a combination of Levetiracetam + Lamotrigine was the most popular polytherapy (6.5%). Additionally, this study determined that polytherapy was associated with greater prevalence of idiopathic/associated symptoms that include anxiety, depression, headache, dizziness, rash, nausea, constipation, and diarrhoea. Lastly, we showcase and discuss the generalisability of these tools to other UK hospital sites and medical conditions, indicating cross-domain EPR-agnostic utility for accelerated clinical and research use cases.

In conclusion, automated information extraction tools applied to routinely collected data can not only be used for service demand monitoring and improvement but to also to enhance epilepsy research. This project unified information across different clinical document types on a scale much larger than previous studies. This has never previously been done; therefore, this project had the ability to sub-group people with epilepsy based upon demographics, clinical manifestations and provide insight into treatment patterns, and optimal patient health outcomes.

Declaration

I, Anthony Shek, confirm that this thesis and the research presented therein is my original work. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. All help received and all published sources consulted are acknowledged appropriately. No part of this thesis has previously been submitted towards any other degree or qualification. Those parts that have been published or submitted for publication are noted as such in the text. All work incorporated in this thesis has been conducted under the supervision of Professor Mark P. Richardson and Professor James T. Teo.

Anthony Chun Shek
28/02/2022

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervision team: Professor Mark Richardson for all of his wisdom shared during our meetings and looking out for my best interests, and Professor James Teo for the amount of guidance, support, motivation, and cooperation that he has selflessly extended, from day one till the very last day, is truly appreciable.

I want to thank the Richardson Lab for their tremendous support and discussions on epilepsy and my peers who contributed, directly or indirectly, to this work, particularly Zeljko Kraljevic, Tom Searle, and Aurelie Mascio who saw me through my first “Hello world” programme to understanding the transformer architecture – but more importantly, all the breaks, lunches, and laughs that we shared made this experience so much more enjoyable.

I would like to thank my partner Alex who has been by my side through thick and thin and knows this thesis down to the last comma (mostly because she corrected them all). Words are not enough to express my gratitude. So, I will just say, thank you for everything you have done for me.

Last but not least, I would like to thank my family – my mother, father, brother and sister for their constant support. The lessons and guidance that you gave me since childhood, has brought me up to this point.

Ethical statement

This project involved working with real-world electronic patient records which contain personally identifiable information (PII) collected as part of routine care. PII is any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another and can be used to deanonymize previously anonymous data is considered PII. This includes but is not limited to names, email addresses, telephone numbers, and national identification numbers.

This project operated under London South East Research Ethics Committee (reference 18/LO/2048) approval granted to the King's Electronic Records Research Interface (KERRI) 2021_05; specific approval in using natural language processing (NLP) on unstructured clinical text for extraction of standardised biomedical codes for patient record structuring was reviewed with expert patient input on a virtual KERRI committee with Caldicott Guardian oversight.

Due to the confidential and sensitive nature of these data several additional steps were taken during handling:

- All data was de-identified.
- No data was transferred or stored on portable devices.
- No data was transported outside the physical boundaries of the respective NHS trust from which the Data was extracted.
- No information was shared or communicated with those who did not have the necessary permissions to access the data themselves.

All required permission for use licences for SNOMED-CT international, SNOMED UK and UK drugs extensions, and UMLS were obtained.

Funding

The King's Medical Research Trust (KMRT) fully funded this work. This included funds for training and development and to present data at conferences.

***

AI will not replace doctors; doctors who use AI will.

Abbreviations

A&E	Accident and Emergency
AED	Anti-epilepsy Drug
AF	Atrial Fibrillation
AI	Artificial Intelligence
AIS	Acute Ischaemic Stroke
ASM	Anti-seizure Medication
BERT	Bidirectional Encoder Representations from Transformers
CCF	Congestive Cardiac Failure
COVID-19	Infectious Disease Caused by the SARS-CoV-2 Virus
CT	Computerised Tomography
CUI	Concept Unique Identifier
CV	Cardiovascular
CVD	Cardiovascular Disease
DSM	Diagnostic and Statistical Manual of Mental Disorders
EEG	Electroencephalogram
EHR	Electronic Health Records
EPR	Electronic Patient Records
ERNIE 2.0	Enhanced Representation through kNowledge IntEgration 2.0
fMRI	Functional Magnetic Resonance Imaging

GDP	Gross Domestic Product
GP	General Practitioner
GPT	Generative Pre-trained Transformer
GSTT	Guy's and St Thomas's Hospital
HES	Hospital Episode Statistics
ICD	International Classification of Diseases
ICT	Information and Communications Technology
ILAE	International League Against Epilepsy
KCH	King's College Hospital
MedCAT	MEDical Concept Annotation Tool
MedCATtrainer	MEDical Concept Annotation Tool trainer
MRI	Magnetic Resonance Imaging
NER	Named Entity Recognition
NHS	National Health Service
NLP	Natural Language Processing
OPCS	Office of Population Censuses and Surveys as the Classification of Surgical Operations and Procedures
PET	Positron-emission Tomography
PII	Personal Identifiable Information
PRUH	Princess Royal University Hospital
QOL	Quality of Life

RNN	Recurrent Neural Network
RoBERTa	A Robustly Optimised BERT Pre-training Approach
SCTID	Standardised Nomenclature of MEDicine Term IDentifier
SNOMED CT	Standardised Nomenclature of MEDicine Clinical Terms
SPECT	Single-photon Emission Computed Tomography
UK	United Kingdom
UMLS	Unified Medical Language System
USA	United States of America
WHO	World Health Organisation
XLNEt	Transformer-XL Model Pre-trained using an Autoregressive Method

Disseminations

This section lists the publications that arose from this thesis.

As first author:

Shek, A., Jiang, Z., Teo, J., Au Yeung, J., Bhalla, A., Richardson, M., & Mah, Y. (2021). *Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing*. European Journal Of Neurology, 28(12), 4090-4097. doi: 10.1111/ene.15071

O'Gallagher, K., Shek, A., Bean, D., Bendayan, R., Papachristidis, A., & Teo, J. et al. (2021). *Pre-existing cardiovascular disease rather than cardiovascular risk factors drives mortality in COVID-19*. BMC Cardiovascular Disorders, 21(1). doi: 10.1186/s12872-021-02137-9

Shek, A., Biondi, A., Ballard, D., Wykes, T., & Simblett, S. (2019). *Technology-based interventions for mental health support after stroke: A systematic review of their acceptability and feasibility*. Neuropsychological Rehabilitation, 31(3), 432-452. doi: 10.1080/09602011.2019.1701501

As contributor:

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., & Bean, D. et al. (2021). *Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit.* Artificial Intelligence In Medicine, 117, 102083. doi: 10.1016/j.artmed.2021.102083

Bean, D., Kraljevic, Z., Searle, T., Bendayan, R., Kevin, O., & Pickles, A. et al. (2020). *Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust.* European Journal Of Heart Failure, 22(6), 967-974. doi: 10.1002/ejhf.1924

Carr, E., Bendayan, R., Bean, D., Stammers, M., Wang, W., & Zhang, H. et al. (2021). *Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study*. BMC Medicine, 19(1). doi: 10.1186/s12916-020-01893-3

Wickstrøm, K., Vitelli, V., Carr, E., Holten, A., Bendayan, R., & Reiner, A. et al. (2021). *Regional performance variation in external validation of four prediction models for severity of COVID-19 at hospital admission: An observational multi-centre cohort study*. PLOS ONE, 16(8), e0255748. doi: 10.1371/journal.pone.0255748

Zakeri, R., Carr, E., Bean, D., O´Gallagher, K., Kraljevic, Z., & Searle, T. et al. (2020). *Biological Responses to COVID-19 Insights from Physiological and Blood Biomarker Profiles*. SSRN Electronic Journal. doi: 10.2139/ssrn.3719886

Nayagam, J., Jeyaraj, R., Mitchell, T., Walder, D., Al-Agil, M., & Shek, A. et al. (2021). *Patterns and prediction of liver injury with persistent cholestasis in survivors of severe SARS-CoV-2 infection*. Journal Of Infection, 82(6), e11-e13. doi: 10.1016/j.jinf.2021.03.029

Ibrahim, Z., Bean, D., Searle, T., Qian, L., Wu, H., & Shek, A. et al. (2022). A *Knowledge Distillation Ensemble Framework for Predicting Short- and Long-Term Hospitalization Outcomes From Electronic Health Records Data*. IEEE Journal Of Biomedical And Health Informatics, 26(1), 423-435. doi: 10.1109/jbhi.2021.3089287

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

In this chapter, we present a broad-level introduction and background to the topics covered in this thesis. We begin by discussing epilepsy and the major contributions of research which has helped to form the current system of classification. We then discuss how patients with epilepsy are managed in resource-limited healthcare environments. Finally, we discuss the importance of information storage, retrieval, and the recent acceleration of artificial intelligence (AI). From which we list the research questions and approach followed in the thesis. As our research is based on a large UK secondary healthcare provider, we try to provide statistics focused on the UK and countries dependent on similar policies to make the thesis more comprehensible and provide relevant context.

1.1 Introduction to Epilepsy

Epilepsy is one of the most common serious chronic neurological conditions affecting men and women of all ages and has a worldwide prevalence (Ngugi et al., 2010). The Global Burden of Disease Study 2016 revealed that there are 45.9 million people who suffer from active epilepsy globally (GBD, 2019). The main characteristic of epilepsy are epileptic seizures, which are defined by the International League Against Epilepsy (ILAE) as “a transient occurrence of signs and/or symptoms due to abnormal, excessive or synchronous neuronal activity in the brain” (Fisher et al., 2014). Epilepsy is commonly defined as a disorder of the brain characterised by an enduring susceptibility to generate epileptic seizures. However, this definition of epilepsy can be difficult to apply to everyday clinical practice, therefore a clinical definition was introduced which describes the presence of epilepsy when any one of the following conditions are met: “(1) At least two unprovoked (or reflex) seizures occurring more than 24h apart; (2) one unprovoked (or reflex) seizure and a probability of further seizures, similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years; (3) diagnosis of an epilepsy syndrome” (Fisher et al., 2014).

The unpredictable nature of seizures limits the patient’s activities and is the primary reason why this neurological disorder can be seriously debilitating, affecting quality of

life (QOL). Common activities such as having a bath, swimming or driving a motor vehicle need to be conducted with extra caution, supervision or even restricted altogether, as an unexpected seizure or loss of consciousness during these activities can potentially be dangerous and, in some cases, fatal.

1.2 Epilepsy as a Neurological Pathology

Epilepsy is not a single disease but rather refers to a class of neurological disorders which cause deleterious functional changes within the brain, specifically excessive and uncontrolled neuronal activity that manifests itself as epileptic seizures. There are many different underlying neuropathological mechanisms that can lead to an epileptic seizure. These include acquired brain injury, neurodegenerative disorders, congenital anomalies, genetic, metabolic, infectious and inflammatory disorders, as well as unknown causes. The cause of persistent seizures is not always immediately obvious and can be difficult to determine – epilepsies of unknown aetiology represent around 50% of all cases globally (Fiest et al., 2017). Other conditions can cause epilepsy by functionally altering a previously normal brain to increase the propensity towards spontaneously generating seizures; this is termed "Epileptogenesis".

The cause of epilepsy should be disentangled from seizure type: seizures can develop from either a focal onset or a generalised onset. Focal onset seizures start in just one part of the brain. The symptomatic manifestation of a focal onset seizure will usually depend on the area of the brain where the excessive neuronal activity originates. These seizures may also “secondarily generalise” or spread to both hemispheres of the brain. In these cases, the focal onset seizure’s symptomatic presentation will often precede the symptoms of a generalised seizure. On the other hand, generalised onset seizures originate and affect both sides of the brain at once.

The effects of seizures on the brain are complex and have to be disentangled from the effects of any primary underlying neurological disease process that has led to increased seizure susceptibility. It is possible that homeostatic changes in a seizure-susceptible brain or even neuropathological alterations may be adaptive and reversible while others are permanent. Identifying common patterns of causes for developing epilepsy and understanding the mechanisms of how seizures are

generated, how they spread and manifest themselves, will help to classify the different types of epilepsy and to develop common management and therapeutic strategies.

1.3 Epilepsy Classification

Epilepsy is a disease with a long history. It was first described in 2,000 BC by the Mesopotamian civilisation as “antasubbu” or “the hand of sin” that was brought onto an individual by the god of the moon (Magiorkinis et al., 2010). The text describes a person whose neck turns left, whose hands and feet are tense and eyes wide open, froth flowing from the mouth, and consciousness being lost.

The modern word “epilepsy” is derived from the Greek verb epilambanein and means “to take hold of” or “to seize” and for many years this was used to connote both the disease and the single seizure (Patel & Moshé, 2020). The term arose from the lack of understanding and influence of magical or religious thinking of that time, when epilepsy was considered an invasion by demons or evil spirits, as divine punishment, and, in general, the intervention of supernatural powers. Unfortunately, this stigma has continued to be attached to epilepsy even up to the present day, leading to many people living with epilepsy feeling compelled to hide their struggle with the disease.

In 500 BC, the foundations for a scientific paradigm shift were set in the book “On the Sacred Disease” from the Hippocratic collection, moving away from the idea that diseases are acts or invasions by gods, demons, or evil spirits, towards rational pathology, stipulating that all diseases arise from merely physiological processes (Hippocrates, 2007). At a time when magicians, wizards, and charlatans perpetuated the popular notion that epilepsy is a spiritual disease, terming it the “sacred” disease, the author disputed the idea by stating ‘*I am about to discuss the disease called “sacred”. It is not, in my opinion, any more divine or sacred than other diseases, but has a natural cause, and its supposed divine origin is due to men’s inexperience and to their wonder at its peculiar character*’. He continues on to consider it as a disease of the brain that was potentially treatable, not through magic, but rather through drugs and diet (Hippocrates, 2007). This turn from a supernatural to a naturalistic explanation of epilepsy is considered a major milestone in the history of medicine.

Nevertheless, for epilepsy it was not until the last 200 years that major progress was made in establishing the classification and terminology for epileptic seizures and syndromes, which provided a fundamental structure for diagnosing, organising, and differentiating this condition. Classification plays a central role in clinical epilepsy and epilepsy research. However, the establishment of a seizure and epilepsy classification system has been a long journey, full of debates on nomenclature, and to this day the system is still continuously evolving.

1.3.1 20th Century Classification Systems

The International League Against Epilepsy (ILAE), founded in 1909, is the largest academic body in the epilepsy community and has been taking a leading role in forming classifications for seizures and epilepsies (Meinardi, 1999). A group of experts, called the ILAE Commission on Classification and Terminology, is responsible for maintaining and updating classification systems regularly.

During the 20th century, the invention of electroencephalogram (EEG), advances in neurosurgery, the discovery of antiepileptic drugs, and the delineation of underlying pathophysiological mechanisms contributed significant progress towards our understanding of epilepsy. By 1964 the ILAE Commission felt compelled to gather and discuss the development of an international classification system because at the time "current classifications of epileptic seizures vary considerably, and the need for a standardised and uniform system of grouping is very apparent" (Commission, 1964). The chairman of the Commission at the time, Henri Gastaut, realised the importance of the relationship between the EEG and clinical seizure semiology and recognised the need for a common method of communication, initiating the effort for the classification of seizures. A few years later, the General Assembly of the ILAE accepted the first publication of clinical and EEG classification of epileptic seizures (Gastaut, 1969, 1970). This classification system was based on a dichotomy of focal (or at the time called "partial") versus generalised seizures. However, this classification was revised several times in the following decades, clarifying the nomenclature and definitions (such as "complex partial") to avoid misunderstandings and inconsistencies during the, at times, challenging practical application of this classification system. In 1981, the ILAE published a revised classification of epilepsies which was based on

the categorisation of seizure types using video-EEG recording evidence (Commission, 1981, 1989).

1.3.2 Modern Classification Systems

At the beginning of the 21st century, the ILAE sought to update the seizure and epilepsy classifications. While it has been clear for some time that updated classifications were needed, consensus was difficult to reach. The classification of the epilepsies was partially extended in 2010 (Berg et al., 2010; Berg & Scheffer, 2011), with the 2010 revision being an intermediate release prior to the final accepted epilepsy classification which would eventually be accepted in 2017.

Given the advances in neuroimaging, genomic technologies, and molecular biology, it was proposed to move towards a classification of seizures based on advanced biomarkers. However, after intense discussions, it was felt that current knowledge of these markers was insufficient to form an exhaustive classification system. Therefore, the updated 2017 classification of seizures continued to rely on semiology, EEG and occasionally supplementary information from imaging (Fisher, Cross, D'Souza, et al., 2017; Fisher, Cross, French, et al., 2017). Interestingly, some of the alternative terminology adopted in the 2017 seizure classification ("focal" instead of "partial", and the term "aware") were terms already considered and debated in 1981, and have continued to stir controversy (Commission, 1981; Falco-Walter et al., 2018). The changes of terminology are represented in Figure 1. Classification of the epilepsy type and epilepsy syndrome does utilise more of these recent advances, being aided significantly by genetics, lab findings and neuroimaging findings, though still with a major focus throughout the diagnostic process on aetiology (Scheffer et al., 2017).

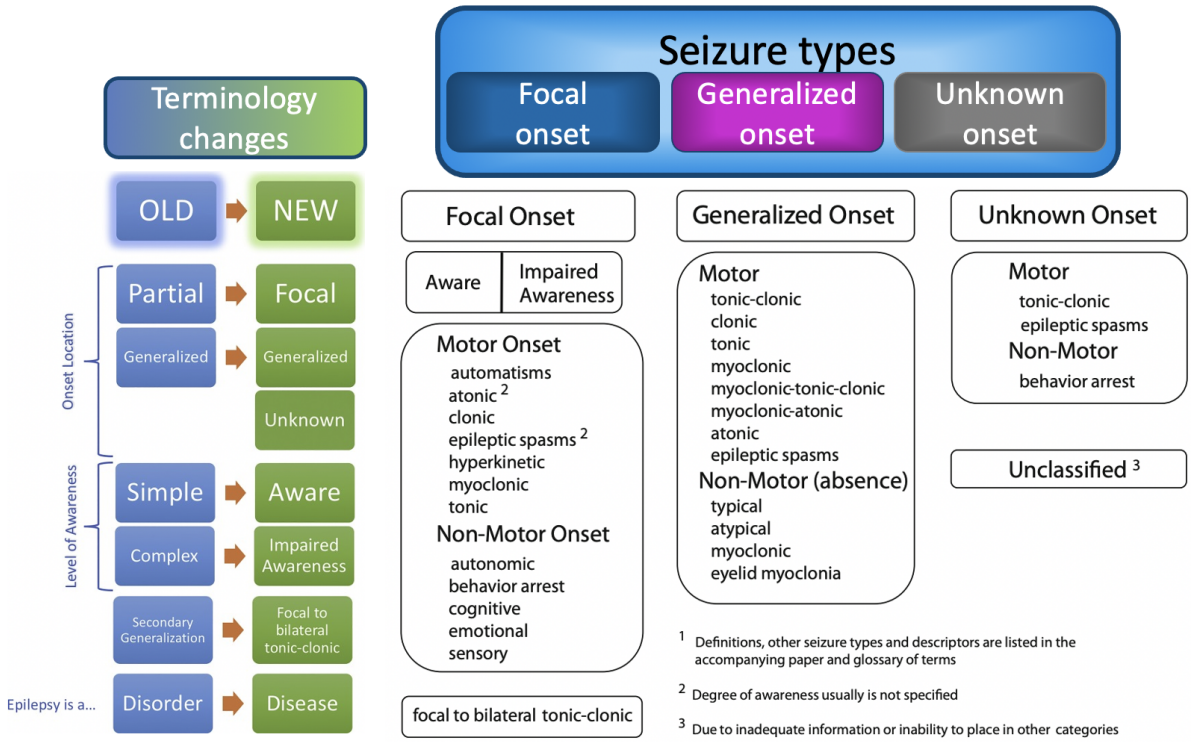


Figure 1. ILAE 2017 Classification system of seizure types and the changes to the terminology (Fisher et al., 2017; Falco-Walter et al., 2018)

The revised 2017 framework supports detailed semiology of seizures as well as the aetiology of epilepsy. It does not represent a fundamental change but allows greater flexibility and transparency in naming seizure types. It was designed to be more easily applied in clinical practice, allowing for vague categories such as seizures of unknown onset (Figure 2). In this classification, the diagnosis of epilepsy always starts with determining the type of seizure, which is then followed by the type of epilepsy. Epilepsy is classified into four main types in this revised classification: 1) focal, 2) generalized, 3) combined generalized and focal, 4) unknown. However, many people, including healthcare professionals, who are not specialised in epileptology may find it difficult to follow the rationales behind the changes in terminology and classification system.

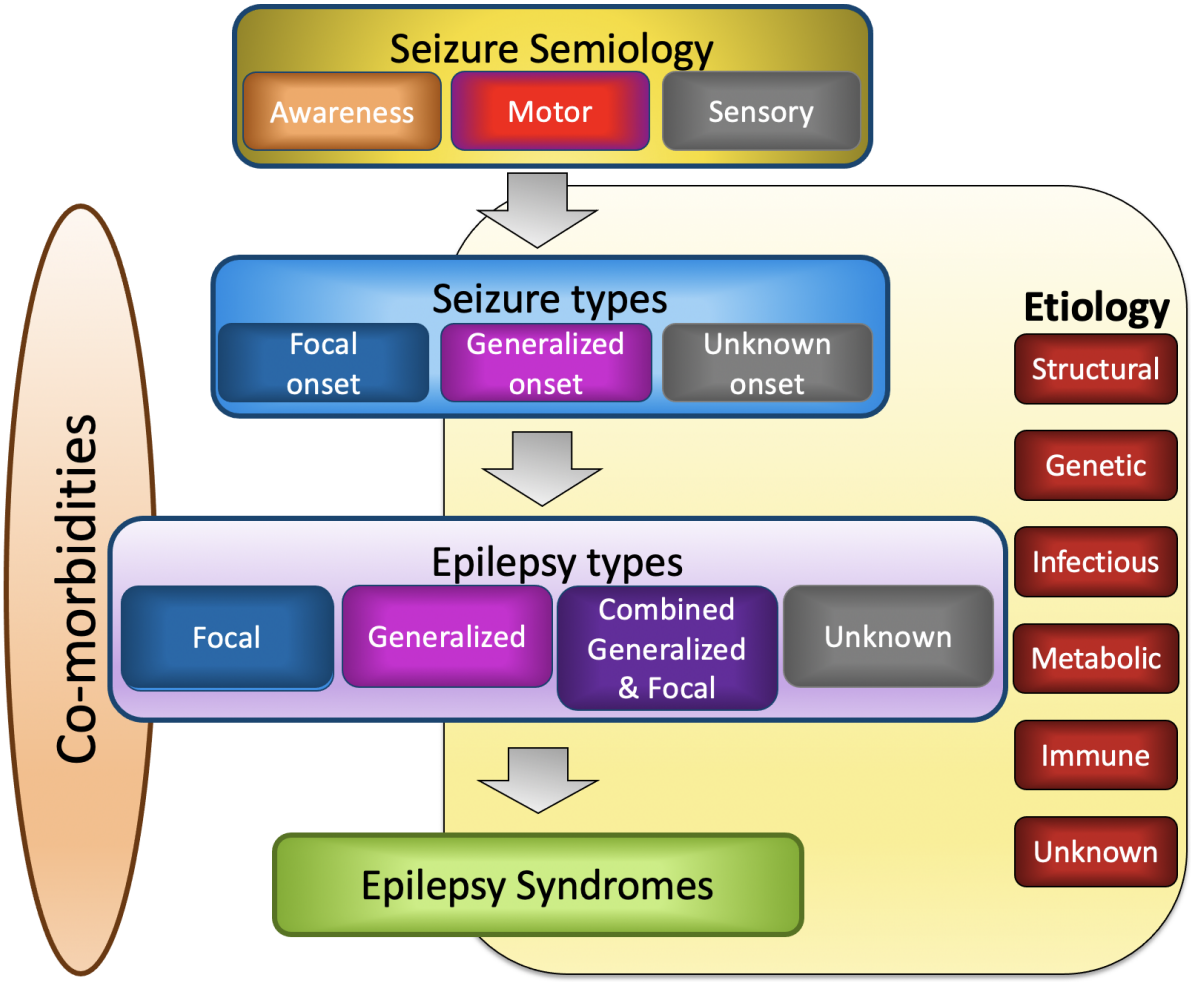


Figure 2. The ILAE 2017 classification of Epilepsy categories modified to include seizure semiology (Scheffer et al., 2017). The directional arrows aim to depict the thought process when determining the most specific epilepsy diagnosis, taking into consideration the aetiology and presence of any comorbidities.

Recently, a tailored classification scheme derived from the 2017 ILAE Classification of Seizures and Epilepsies has been proposed specifically for neonates (Pressler et al., 2021). This is because the 2017 classification has been primarily developed for older children and adults, and is not suitable for neonates. For example, it is common for neonates to have seizures which are only EEG detectable with no evident clinical features (Mizrahi, 1987; Mizrahi & Kellaway, 1987).

The current 2017 ILAE Classification of Seizures and Epilepsies is by no means complete, as evident from the expansions to suit different age groups such as neonates, and will continue to be adapted as our understanding of seizures and epilepsy grows. However, the rapidly changing epilepsy classification system is confusing from a patient perspective. It is challenging for clinicians without an epilepsy expertise to keep up to date with the latest terminology changes, let alone for patients

or the general public. This impedes epilepsy patient-clinician communication and patients' own awareness of their condition and health management. A stable, universally accepted classification system is important as it impacts the communication of epilepsy in clinical care, teaching, and research.

1.4 Epilepsy Pathways

1.4.1 Pathways – First Seizure

Seizures are a common presentation to hospitals in England with up to 100,000 admissions per year. Twenty-two percent of these are first seizure cases and therefore present a considerable burden to emergency departments (Dixon et al., 2015).

If a person suffers from a suspected first seizure event, they initially present either to their GP ('I had a seizure recently') or directly to the emergency department. If after screening an epileptic seizure is suspected or if there is diagnostic doubt, the patient should be referred to a first seizure pathway, where investigations, diagnosis, and initial treatment decision are made by a specialist (who is usually, but not always, a neurologist – in some cases it might also be a general paediatrician or another specialist with particular epilepsy expertise) (NICE, 2021a).

By far the most useful aid for an epilepsy diagnosis is the patient's clinical history, a good description of the events from a witness, or as is commonly the case, a mobile phone video of the events taken by a bystander. The specialist would try to establish the epilepsy type, seizure type, and underlying cause from a combination of clinical history, physical examination, and investigations; besides clinical history, EEG and brain scan (usually MRI) are often crucial in this process (Krumholz et al., 2015).

An EEG is used to look for the presence of abnormal neural activity which may remain after a seizure (Noachtar & Rémi, 2009). People with some types of epilepsy were found to have unusual electrical brain activity even when they are not having a seizure. The results from an EEG investigation can provide supportive diagnostic evidence in the form of specific epileptiform activity patterns that are associated with certain types of epilepsy (Benbadis et al., 2020). However, many people with epilepsy only have unusual electrical brain activity while having a seizure. The rest of the time, the brain

activity may be entirely normal. So, if an EEG test does not show any unusual activity, it only means that there is no epileptic activity in the brain at the time the test is being done. Consequently, this does not rule out the presence of abnormal activity in the brain at other times nor excludes a diagnosis of epilepsy. Nevertheless, the timing of an EEG after a first seizure event is important as it has been shown that early investigations are associated with a higher diagnostic yield (Llauradó et al., 2020).

Structural brain imaging techniques such as MRI or CT scans are also important tools, used to complement EEGs to look for any anatomical abnormalities and potential causes of seizures, e.g. tumours or abnormal blood vessels which can potentially be treated surgically (Duncan, 2019). Functional brain imaging techniques such as fMRI, Positron-emission tomography (PET) and single-photon emission computed tomography (SPECT) may assist in identifying focal functional abnormalities that help infer the location of an epileptic focus. However, with these tools, absence of aetiological features does not rule out a diagnosis of epilepsy but rather the presence of features will provide further evidence for a diagnosis of epilepsy and even shed light on subtypes or syndromes.

Overall, epilepsy remains primarily a clinical diagnosis and the timely management of first seizures is crucial to the subsequent prognosis of the patient and essential for developing an effective treatment strategy.

1.4.2 Pathways – Chronic Epilepsy

Once an epilepsy diagnosis has been made, the next step is to establish the specific details of the patient's condition, as epilepsy can vary from patient to patient in the type, severity and frequency of the seizures. In the UK, there are currently no formally defined care pathways which standardise the management of chronic epilepsy. However, the general guidance is that seizure type(s) and epilepsy syndrome, aetiology, and any confounding co-morbidities should be determined, because failure to classify the epilepsy syndrome correctly can lead to inappropriate treatment and persistence of seizures (NICE, 2021a).

Investigations occur with an epilepsy specialist, most often in a secondary care setting with EEG and neuroimaging facilities – primarily to improve outcomes for patients by ensuring everyone is investigated properly and has the optimal therapy decisions. Once a patient has reached a stable well-controlled epilepsy, they should be managed entirely in primary care settings – ideally this should be the majority of people who have received an epilepsy diagnosis.

High quality management provided through healthcare services is critical for epilepsy patients from the point of diagnosis. This is because epilepsy cannot be cured and patients can have frequent interactions with a healthcare provider. Patient healthcare service provision and management are often intertwined and highly correlated with the patient's health trajectory. The inability to get timely medical appointments, a delayed diagnosis, suboptimal treatment strategies, or poor patient self-care and lack of understanding of their own condition will detrimentally affect their health trajectory and in turn their quality of life. This can increase the long-term cost to the healthcare system and the overall amount of resources required to manage these patients.

1.5 Anti-Seizure Medications

Achieving seizure-freedom is the primary goal of the treatment and management of epilepsy. Anti-seizure medications (ASMs), formerly known as anti-epilepsy drugs (AEDs), are the first line of treatment for new-onset epilepsy. There is a wide variety of ASMs available that have different targets and mechanisms of action in the brain. This choice is based on the specific needs of the patient, including factors such as age, lifestyle, other health problems, and the category or subcategory of epilepsy. To date, the best source of evidence for the optimal choice of ASMs are the SANAD and SANAD II clinical trials, comparing the longer-term effects of ASM treatments (A. Marson et al., 2021a, 2021b; A. G. Marson, Al-Kharusi, et al., 2007a, 2007b).

However, despite there currently being over 20 ASMs, 30% of people with epilepsy do not respond to first line ASMs, and the lack of response is difficult to predict (WHO, 2006). The options for patients who do not respond well to their first ASM include altering the dosage, switching ASM and/or trying them in different combinations. Most patients with refractory epilepsy, meaning that their seizures are difficult to control,

combine 2, 3 or even 4 ASMs to control their seizures. However, the number of possible permutations of ASM combinations are impossible for patients to try within their lifetime. Additionally, there are no guidelines and limited to no evidence for the optimal order of ASM switching or combination for people with epilepsy to follow because clinical studies on ASM combinations are insufficient and difficult to perform (Verrotti et al., 2020).

Importantly, polypharmacy has been associated with worse quality of life (QOL) in patients with epilepsy despite seizure control (Alexander et al., 2018). It increases the potential for negative medication effects and drug interactions, amongst which neurobehavioural adverse effects and behavioural disruption are common. The practice of initiating new medications to manage the adverse effects of other agents, also coined the “prescribing cascade effect” (Rochon & Gurwitz, 1997), is common, with physicians sometimes falsely attributing behavioural disruption to other factors, including the epilepsy itself (Plevin et al., 2018). We do know that many of these drugs have differing pharmacokinetic properties, meaning how our gut, liver and kidneys metabolise them, leading to many different possible side effects, especially when combining multiple drugs.

Despite the fact that all ASMs have been developed to control seizures, seizures may be just one phenotypic expression of epilepsy. Adopting a more holistic view of epilepsy, we know that other neurologic pathways such as the serotonergic and dopaminergic pathways in the brain are also affected by the disease. These neurological pathways are responsible for maintaining the normal functioning of the brain such as executive function or impulse control, areas of the brain that can affect mood. From this, it is not surprising that mood disorders and epilepsy often go hand in hand. We know that some ASMs may in fact impact mood disorders, in some cases making these conditions better whereas in other cases making them worse (Gibbons et al., 2009; Ketter et al., 1999; Muzina et al., 2005).

The key challenge is understanding the ASM choices and combinations as well as how to tailor them to individual patients to achieve the right balance between seizure control, negative drug side effects, and treating other symptoms of epilepsy. ASM therapy requires timely and regular administration, often over long periods of time,

ideally with no or minimally experienced side effects as possible which can include allergic reactions, dose-related effects, and idiosyncratic side effects; some of these can be short term effects vs others remaining prevalent in the long term.

Ensuring lasting success of seizure freedom poses further challenges: even if the ASM therapy is successful in restoring a patient's control over their seizures, this does not always mean that they should stop taking their medications as there is always a risk that seizures may reoccur if these medications are ceased (Strozzi et al., 2015). This makes weaning patients off ASMs extremely complicated.

Newly developed ASMs show reduced risk of adverse events and have unique or different mechanisms of action that enable the creation of possible synergistic combinations (Kwan & Brodie, 2006; E. Perucca et al., 2007). Unfortunately, in terms of seizure freedom efficacy, they display similar results to traditional ASMs (A. G. Marson, Appleton, et al., 2007; Rowan et al., 2005). If seizures are unable to be controlled with medication, other treatment options include epilepsy surgery where specific brain areas that are thought to cause the seizures are surgically removed, for example a tumour. Nerve stimulation might also be considered where certain nerves such as the vagus nerve are stimulated, which is thought to control seizures by influencing neurotransmitter release (Ben-Menachem, 2002). Also for specific epilepsy syndromes, patients might adopt a ketogenic diet, reducing carbohydrates from one’s diet to force the body to burn fat instead of carbohydrates, producing ketone bodies which are then used by the brain as an energy source instead of glucose (Y.-Q. Wang et al., 2020). Although many hypotheses have been suggested for why this seems to reduce seizures for some patients, the exact mechanism is not well understood.

Overall, it is important to do what we can as soon as we can to achieve the best outcome for patients. Scientific research is aiming to provide clinicians with every tool and every insight available to easily monitor people with epilepsy and to allow them to rapidly develop a management plan. A key focus should be to facilitate the choice of ASM or combinations thereof to optimally match patient needs and allow for evidence-based drug-switching or mixing to enhance patient outcome over time.

1.6 The Modern Healthcare System and its Challenges

1.6.1 Electronic Patient Records

Hospitals generate data daily, ranging from patient statements to reports by medical professionals, from diagnostic tests to appointment booking systems. The increasing adoption of electronic patient record (EPR) systems allows both the collection and storage of information in a digital format. EPR systems have primarily been developed for direct patient care, to accurately capture the state of the patient across time. The NHS deals with over 1 million patients every 36 hours (Department of Health, 2005). Therefore, each NHS trust can be seen as a health data factory, holding a wealth of patient information including demographics, medical history, medications, and test results.

Pooling patient information into data registries may help reveal population-level data not readily apparent at the individual level. To do this, EPRs can extract discrete data from administrative and clinical sources to create a “data warehouse” (MacKenzie et al., 2012). The information architecture underlying a data warehouse is designed to support queries and analysis of populations; this is different from the architecture of clinical EPR systems designed to rapidly access individual patient information (Lyman et al., 2008). In an increasingly digital age for healthcare around the world, electronic patient health records data have become rich and accessible tools for identifying and monitoring healthcare use of both individuals and whole populations. Taking a holistic view of the healthcare system, this data could also help improve the delivery of healthcare by means of cost savings, improved safety, and increased operational efficiency (Mbizvo et al., 2018; Mbwana et al., 2019).

Epilepsy is a good example to demonstrate how maximising the utility of data contained in EPR systems could improve the healthcare system. People with epilepsy often have frequent encounters with healthcare providers and the condition itself is considered to be ambulatory care sensitive, meaning that high quality care may reduce unnecessary future inpatient and emergency department follow-up (Grinspan et al., 2018; Wilner et al., 2014). Since each interaction is recorded, the information contained in EPR can help to identify patterns of service utilisation and any

discrepancies provided to patients early on, potentially reducing the long-term cost to the healthcare system. This is especially important when considering that the main driver of the cost of care for people with epilepsy is hospital service utilisation (Taylor et al., 2011). This substantially increases when co-occurrence of other medical conditions are present as patients are at increased risk of readmission and generally incur medical costs almost 1.4 times higher when compared to people with epilepsy without such comorbidities (Lee et al., 2005). Overall, EPRs can not only act as a data source for patients' past medical history, services provided, and interventions in an ecological valid environment but can be used to learn from previous patients' experiences to make better informed outcome predictions about treatment and possible health trajectories for new incoming patients with similar demographic and health profiles, to increase their chance of seizure freedom and quality of life.

1.6.2 Standardised Medical Language Classifications and Terminologies

An important consideration is that each hospital's electronic patient record system and any interoperable national health information infrastructure requires the use of uniform health informatic standards (Ayaz et al., 2021). This includes the utilisation of a common medical language to facilitate effective and efficient communication between healthcare industry staff. Data must be collected and maintained in a standardised format, using uniform definitions which are agreed upon and universally accepted, in order to link data within an EPR system or share health information between hospital systems. The lack of standards has been a key barrier to electronic connectivity in healthcare (McDonald, 1997). Standardised clinical classifications and terminology systems represent a common medical language, playing different but critical roles in allowing clinical data to be utilised and shared within and between EPR systems.

Classification Systems

A classification is a system for organising knowledge about the similarities among and differences between items that are part of some overarching group. While diagnoses can be organised within a classification structure, the classification structure is not essential to the diagnosis or even the thought process in determining a diagnosis. A diagnosis is the identified disease or disorder for a particular individual that is arrived at through the process of considering history, signs, symptoms, and other clinical

information. In epilepsy, the term ‘‘classification’’ is also often used to refer to the list of the different forms of epilepsy and to an individual diagnosis itself (e.g., ‘‘The Patient's epilepsy was classified as childhood absence epilepsy’’).

Classification systems such as DSM, ICD, OPCS, group together similar diseases and or procedures and organise related entities for easy retrieval (American Psychiatric Association, 2013; NHS Connecting for Health et al., 2006; World Health Organization, 2004). They are typically used for external reporting requirements or other uses where data aggregation is advantageous, such as measuring the quality of care, monitoring resource utilisation, or processing claims for reimbursement. Classification systems are considered “output” rather than “input” systems and are not intended or designed for the primary documentation of clinical care. They are inadequate in a reference terminology role because they lack granularity and fail to define individual clinical concepts and their relationships. Yet, they are the most common source of clinical data today, readily available as a by-product of the healthcare reimbursement process.

Terminologies

Terminologies are inadequate for serving the secondary purposes for which classification systems are used because of their immense size, considerable granularity, complex hierarchies, and lack of reporting rules. Reference terminologies such as RX-NORM, HPO, LOINC, UMLS, SNOMED-CT are primarily used for direct structured medical reporting or inferring and structuring medical knowledge from clinical text (Bodenreider, 2004; Köhler et al., 2019; Liu et al., 2005; Maloney & LOINC Developers, 2003; *SNOMED CT Starter Guide*, 2021). The benefits of usage increase exponentially if the reference terminology is linked to modern, standard classification systems for the purpose of generating health information necessary for secondary uses such as statistical and epidemiological analyses, external reporting requirements, measuring quality of care, monitoring resource utilisation, and processing claims for reimbursement. The linkage of terms within and between different systems to extract information for multiple purposes is accomplished through mapping. These can be flat (siblings) or hierarchical (parents and children) in nature, with the latter being useful for grouping concepts under a single representative category.

Overall, neither a clinical terminology nor a classification can, by itself, serve all of the purposes for which health information is currently used or will be used in the future. Terminologies and classifications are designed for distinctly different purposes and satisfy diverse user data requirements. However, together, standard clinical terminologies and classifications represent a common medical language that allows clinical data to be shared between EPR systems. Therefore, standard clinical terminologies and classifications, with rule-based maps linking them, must be incorporated into EPR systems in order to achieve system interoperability and the benefits of a national health information infrastructure. Nevertheless, the full value of the health information contained in an EPR system will only be realised if the mapping between the classification and terminology systems is up-to-date and accurately reflects the current practice of medicine.

1.6.3 Value-based Healthcare System

Today, we still see dysfunction in healthcare systems around the world despite an exponential increase in medical knowledge. Estimates indicate that a fifth of healthcare spent is wasted (OECD, 2017). There is wide variation in both health outcomes and costs across UK providers within the NHS (Rodriguez Santana et al., 2020), let alone between countries (Papanicolas et al., 2018). Currently, costs are rising at an unsustainable rate outpacing GDP, despite attempts from governments and other funders to control costs (OECD, 2017). Additionally, rising public expectations about the quality of healthcare and the rapidly expanding development of new technology-based medical facilities, increase upward pressures on healthcare spending. We need a healthcare system designed in a manner where patients and stakeholders are aligned on the definition of success, that is delivering the best possible outcomes for patients at the lowest cost. This is known as a value-based healthcare system. An increase in value can not only come from increasing patient outcomes at the same cost, but also from decreasing the cost of resources required without sacrificing patient outcomes.

$$Value = \frac{Patient\ Outcome}{Cost/Resource\ allocation}$$

The ideology of ‘value’ is gaining prominence in the healthcare industry as it faces an ever-increasing demand for services with limited resources. Value-based healthcare emphasises the value received from interventions or services provided in proportion to their respective cost (Gray, 2017). Faced with limited resources, simply increasing spending on healthcare to improve services and outcomes may not always be possible, instead optimal resource allocation requires an informed calculation based on information about the opportunity cost of the inputs to healthcare as well as the value of the benefits received. Similarly, value-based medicine incorporates the best features of evidence-based medicine (ensuring that only interventions with strong evidence are used) and considers evidence-based cost data. However, it is not suggested that this train of thought should be used by healthcare professionals directly interacting with patients under their care but rather by the designers and operators of the healthcare service as a whole to help them continue to provide the highest quality services in an economically sustainable manner.

The main challenge of this school of thought is that the heterogeneity of patient outcomes are so numerous that they can be difficult to feasibly collect, aggregate, and be used for comparison at scale. Current measures of outcomes are usually constrained to the improvement gained in length of life, but generally ignore the improvement or loss of quality of life. Ideally, outcomes should be broken down or aggregated into different levels of granularity e.g. specific test results which are already stored within the electronic patient record (EPR).

It is therefore virtually impossible for any one study to undertake a comprehensive and unambiguous assessment of the total returns to health care investment, including all sources of benefits and negative consequences. Nevertheless, there is scope for considerable improvement in monitoring and evaluating health spending so as to help better inform future decision making. The development of evidence-based practice and the establishment of institutions such as National Institute for Health and Clinical Excellence (NICE) are attempts to begin to evaluate and regulate potential new health care costs. One way they do this is through setting guidelines to standardise and optimise the provision of UK healthcare patient care pathways. Yet, such institutional innovations still only address a subset of the problems faced by the NHS.

The UK healthcare system is intrinsically value-based in its design since there is a fixed capacity of funding for delivery of care, however this relies on efficient data flows to match capacity with demand and cost-effectiveness. However, the combination of complex administrative or competition rules (which produce perverse incentives), and legacy technology systems, results in a lack of realisation of 'value'. A focus on 'value' can transform the current healthcare system into a more sustainable, patient-centric system supported by evidence-based decision making. This starts with measuring health outcomes and combining informatics with key enablers, facilitated by a supportive public policy environment (Majeed et al., 2018; Robertson et al., 2010; Sheikh et al., 2011). Through optimising the provisions of the healthcare service, meaning the ability to supply the correct diagnostic tests or interventions at the right time during a patient's disease trajectory, this will increase the probability of positive health outcomes. Together, enterprise scale monitoring of health outcomes and hospital service utilisation is becoming more feasible given the adoption of electronic patient records, as they record interactions between healthcare providers and patients. The goal would be to create an evolutionary healthcare ecosystem where outcome transparency allows us to learn from systems across different healthcare providers, identifying best practises and eliminating wasteful spending that does not drive value.

1.6.4 Digital Infrastructure Barriers to Clinically-actionable Analytics

The utilisation of data contained in EPR for secondary purposes such as research or service development is not new. However, to date, large scale applications still face several key challenges to the retrieval of information. Quintessentially, it is a "big data" problem. Despite the fact that hospitals still hold the vast majority of data captured in EPR, there is a huge amount of clinically relevant information fragmented across various documents, e.g. physician clinic letters, lab results, images which are contained across different document formats, including pdfs, doc, plain text, etc. These are generated rapidly at a speed much faster than it is possible and feasible to analyse and act on them. EPR systems are closed, proprietary, and often contain incomplete information. Additionally, there is an inadequate integration of intra- and inter-healthcare systems and poor healthcare information management (Y. C. Wang et al., 2018). Disjointed legacy information and communications technology (ICT)

infrastructure and different EPR system providers at each hospital are the primary causes of this. This is a problem as patients may attend multiple hospitals during ongoing investigations and treatments. They may seek second opinions from doctors at different hospitals, have an unexpected health incidence whilst travelling, or move out of hospital catchment areas for other reasons such as employment opportunities or retirement.

1.7 Information Extraction and the Recent Acceleration of AI Innovation

1.7.1 Automated Information Extraction from Textual Data

Textual data can be presented in two ways: structured and unstructured. Structured data refers to data input that is stored into a predefined format. Examples of this are age, gender, codes or definitions from terminologies or classification systems. These data can be directly processed and analysed as the input and format are known. However, the majority of data held in EPR systems are unstructured, including clinical notes, discharge summary notifications, diagnostic investigation reports, and clinic letters from hospitals and General Practitioner (GP).

Given that at least around 80% of all health data are unstructured, the ability to accurately analyse unstructured text plays a pivotal role in the success of big data analytics in healthcare (Assale et al., 2019; Wang et al., 2018). Unstructured text in EPR is ubiquitous as it is used to keep track of the health of patients and serves as a more natural and expressive way to document clinical events and facilitate communication among the care team in the healthcare environment. For this reason, unstructured text has the ability to provide complete descriptions of reported health symptoms, interventions provided and response, whilst maintaining high ecological validity. Unstructured and semi-structured data in healthcare refer to information that can neither be stored in traditional relational databases nor fit directly into predefined data models (Assale et al., 2019).

It is also well known that unstructured data contains relevant, richly detailed, and nuanced information about illness trajectory and care processes undertaken by and upon patients (Vest et al., 2017). This information can be manually extracted from

EPR; however, this is a tremendously time-consuming task requiring expert knowledge, making the challenge to automatically extract accurate information from unstructured texts worth pursuing (Carrell et al., 2017).

The ultimate goal of information extraction is to automate, i.e. to have computers 'read' document contents to extract and organise any relevant information. In other words, the goal of information extraction is to convert unstructured text into a structured format subject to the user's needs. With the deluge of big data and new avenues for collecting information, structured data is no longer the only source from which meaningful information can be derived on a large scale. With advances in computer power and infrastructure that allow users to cope with growing volume, velocity, and variety of datasets, it has become more feasible to analyse unstructured data sources in the healthcare system (Assale et al., 2019).

Overall, EPR information, particularly unstructured data, has the ability to provide insights beyond direct patient care and into diseases, associated symptoms, treatment strategies, adverse events, and healthcare service usage. Together, automated information extraction can improve our knowledge of not just diseases and their prevalence, but also optimise patient management and service usage through evidence-based decisions.

1.7.2 Natural Language Processing and Recent Developments

Unfortunately, automated information extraction is not a straightforward task. The Prussian philosopher and linguist Wilhelm von Humboldt described language as a system that involves the "infinite use of finite means" and this quote perfectly encapsulates the nature of unstructured text. Natural language processing (NLP), defined as the use of "computational techniques for the automatic analysis and representation of human language" (Chowdhury, 2003) has been attempting to tackle this problem through automation. NLP approaches to information extraction can be broadly classified into two categories: rule-based and machine learning, with the latter recently increasing in popularity to tackle more complex tasks.

Rule-based NLP techniques are conducted through a sequence of characters that specifies a search pattern. Usually, such defined patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. These have been enormously helpful during the writing and editing of this thesis and I am sure to the reader skipping through to different sections! These can be fast to write and run and easy to troubleshoot, however, these rules can become convoluted very quickly especially during complex queries.

On the contrary, in machine learning-based information extraction techniques, extraction rules are neither explicit nor defined but rather ‘learnt’ through providing the computer with ‘correct’ examples. Since computers cannot ‘read’, a popular NLP technique is to represent each word or phrase within a corpus as a different high dimensional mathematical vector known as ‘word embedding’. These embeddings aim to capture syntactic patterns which can be used to infer meaning and are fine-tuned or ‘learnt’ through a neural network. This method of training a computer system is modelled after the human brain, with ‘hidden’ processing layers between the inputs and outputs, which is why the technique is often lumped into the field of deep learning. Due to the accelerating availability of computational power, neural network-based algorithms have evolved rapidly over recent years, becoming more commonly used in NLP.

In 2013, Google released Word2vec which was based on the idea that neural networks can learn similarities between words based on word distributions and represent each unique word with a corresponding vector (Mikolov et al., 2013). For example, the word “fit” can be used interchangeably or near similar words such as “healthy” or “exercise”, and therefore it should be mapped close to them in the vector space. This creates a graph of all words within a vocabulary where their respective vector representation (embedding) encodes its meaning, such that the words that are closer in the vector space are expected to be similar in meaning. However, from our example, “fit” can also be used in different contexts to represent an alternative meaning such as “seizure”, “healthy”, or to describe other forms of attacks. This highlights the limitation of this technique by the fact that it does not disambiguate the meaning of homonyms contextually.

The 2014-2016 period was marked by the rising popularity of the recurrent neural network (RNN). The idea of an RNN is simply that all weights/parameters are reused for each iteration step. Every hidden state is updated depending on the previous hidden state and the current input. The hidden state works as the memory of the network that stores the most important information from the previous inputs, thus overcoming the limitations of Word2vec to be able to disambiguate homonyms based upon the context in which they are used. One problem with this method initially was that although it worked well for short sentences like “I am working hard”, which can perhaps be captured by a single embedding, it did not work well for long and complex sentences like “While I was working, the telephone was ringing”. This is because as the RNN block reached the end of the sequence, it was ‘overwhelmed’ by the input length and ‘forgot’ what the beginning of the sentence was. This problem of ‘forgetting’ was eventually solved by a mechanism called ‘attention’ (Bahdanau et al., 2014). This ‘attention’ mechanism is nothing more than a way to decide which part of the sentence should receive more importance. In other words, attention tells an RNN what it should focus on or what not to forget. Yet another problem with recurrent neural networks is that sentences are processed sequentially, word by word. However, for a machine reading line by line means it needs a lot of time to process a huge corpora of text.

To overcome that problem, the transformers architecture was proposed in the paper “Attention is all you need” (Shazeer et al., 2017). Transformers, like RNNs, handle sequential data, but unlike RNNs they are not required to do so in a particular order. This means that they can be trained much faster with parallel processing, i.e reading multiple sentences at once. In 2018, Google researchers developed a bidirectional version of the transformer which they called bidirectional encoder representations from transformers, aka “BERT” (Devlin et al., 2018). The key term here is bidirectional, allowing information to flow forwards and backwards as the model trains, resulting in better model performance and language representation. Similar to Word2vec, BERT also maps similar words more closely, with the advantage of being context-sensitive whilst, as mentioned previously, Word2vec is context-independent. This means that BERT computes a different word embedding for a word if it is encountered in different contexts. Reusing the previous example, if we had two sentences “this patient is fit and well” and “this patient with epilepsy had an epileptic fit”, the word fit will have two

very different word embeddings stored within BERT, arising from the two different contexts, allowing it to disambiguate words contextually. This addresses the weakness of Word2Vec for homonyms. Nevertheless, this is not the end of the NLP evolution; researchers developed RoBERTa, ERNIE 2.0, XLNET, and GPT, all of which extend the core transformer framework, both conceptually and in performance. New architectures rapidly continue to be developed and released to the present day.

Although natural language generation is not explored within the scope of this thesis, it highlights the current capabilities and future potentials of the latest NLP models to capture linguistic patterns from a corpus of text. As a demonstration, the following paragraph has been generated using a model called the generative pre-trained transformer-2 (GPT-2). GPT-2, is a 1.5B parameter transformer model trained on a dataset of millions of web pages, called WebText, with a simple objective: predict the next word, given all of the previous words within some input text (Radford et al., 2019). No further fine-tuning was conducted on this pretrained model. The input for this demonstration was all previous paragraphs from this thesis' introduction section with the addition of the phrase "Aims and objectives: ". The model was then required to generate a paragraph of max length 150 words:

> "*The aims and objectives of this thesis is to develop a framework to support the identification of medical information within unstructured medical text using advanced machine learning techniques and to provide a more robust approach to healthcare information analysis and discovery. The framework will: (i) provide a model to identify and extract medical information from unstructured medical text; (ii) describe the design of future advanced machine learning systems that can reliably identify and extract medical information, (iii) explain how these new systems will help clinicians make better informed decisions about the management of patients with epilepsy; (iv) describe the potential and limitations of these new systems for clinical decision-making and healthcare information discovery .*" – This section of text was auto-generated text using the GPT-2 pipeline from the HuggingFace transformer software library (Wolf et al., 2020).

Turning back to natural language extraction, it is important to measure an algorithm's ability to extract meaning out of text and to assess its accuracy. To do so, the performance of information extraction NLP models are commonly benchmarked against a human-created labelled dataset where the labels are assumed to be the

ground truth (how a human would extract information vs an automated system) and the performance can be measured quantitatively. The quantitative measures are a) precision (aka positive predictive value, a ratio of relevant instances among all retrieved instances), b) recall (aka sensitivity, a ratio of retrieved instances among all relevant instances), and c) F-score (aka harmonic mean of the precision and recall). The following are their formulae:

A. $Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$

B. $Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$

C. $F-Score = \frac{2(Precision)(Recall)}{(Precision+Recall)}$

When it comes to comparing information extraction models, it is not as straightforward as comparing their F-score. This is because the comparison of models remains subjective, depending on the task at hand, how the train-test split has been conducted, and which datasets they have been evaluated on.

Overall, advancements in NLP systems, particularly those designed for information extraction tasks, provide an alternative to using highly structured forms or to directly inputting data into “data warehouses”. Rule-based and/or machine learning-based techniques have the potential to mine and extract clinically meaningful information from unstructured text. Yet, artificial intelligence should not be considered a silver bullet but rather a set of tools that can help us improve our understanding and provision of healthcare, as it is still far from perfect. We still require standardised and generalisable NLP performance benchmarks to even begin to answer the question of how accurate is accurate enough? Nevertheless, the UK government recognises the vast potential of AI to improve patient care pathways and the way we administer healthcare services and has outlined its ambitions through a strategy to incorporate future AI-driven technologies into its national health service (NHS) (NHSX, 2021).

1.8 Objective, Aims and Scope of this Thesis

This thesis will provide a fresh insight into correlations between patient care pathways and specific outcomes of patients with epilepsy, at a scale not previously possible with traditional study designs or manually created data sources. This will be done by centralising, capturing and aggregating large amounts of structured and unstructured clinical narratives from heterogeneous textual data sources, i.e. different types of medical records and administrative records from a secondary care setting – King's College Hospital NHS Foundation Trust (KCH), using an information retrieval and extraction platform called CogStack. These records will be grouped, and any relevant information within the records will be automatically extracted using natural language processing, then analysed with descriptive analytics.

The aims of this thesis are:

1) Develop an accurate tool for extracting relevant clinical information from unstructured clinical records at scale.

2) Identify key patient features of two cohorts: patients who have a first suspected seizure case in 2017; and patients with an epilepsy diagnosis who have attended an outpatient neurology clinic between 2013-2020.

3) Stratify patients based upon patient-level features such as treatment responsiveness, associated symptoms, comorbidities and healthcare service utilisation.

4) Map and explore the relationships between people with epilepsy's health trajectories, anti-seizure medication (ASM) choice and hospital services usage.

5) Demonstrate the generalisability of this approach beyond epilepsy to other diseases and beyond a single NHS Hospital.

Chapter 2: Methods

In this section, we will discuss the methods and tools that were used during this thesis and the rationales why they were used:

1. The CogStack ecosystem was used to centralise and uniformly format electronic patient records (EPR) and associated meta-data.
2. Python version 3.6+ together with python version compatible libraries was used for processing, analysis and visualisation of the data.
3. Systematized nomenclature of medicine clinical terms (SNOMED CT) was used to standardise the information retrieval or extraction through unique concept representations of clinical knowledge.

In summary, these tools were chosen because they were readily available, set up and adopted by their respective fields prior to the beginning of this PhD, and more importantly they were the most suitable tools for the information extraction task at hand.

2.1 The CogStack Ecosystem

2.1.1 What is CogStack?

CogStack is an open source, lightweight distributed, fault tolerant database processing architecture, intended to harmonise, standardise, and centralise documents and information found within and across real-world EPR systems (Jackson et al., 2018). It is EPR vendor agnostic and opens up the potential for useful clinical microservice applications to connect and use this framework as a data source. It provides a configurable data processing pipeline, which for the moment mainly uses databases and files as the primary source of EPR data with the possibility of expanding through the addition of custom data connectors for downstream applications in the near future (*CogStack Documentation - Confluence*, 2021).

2.1.2 The Core CogStack Technology Stack

Apache Tika

The Apache Tika toolkit[1] detects and extracts metadata and text from over a thousand different file types (such as doc, PPT, XLS, and PDF). This is extremely useful as EPR documents are stored across hundreds of different file formats.

Tesseract

Tesseract[2] is an optical character reader (OCR) software. OCR systems transform a two-dimensional image of text that could contain scanned material or handwritten text from its image representation into machine-readable text. The purpose of including this technology is to process the ingestion of scanned images of documents found within EPR systems.

Java Spring Batch Framework

It makes use of the Java Spring Batch framework[3] in order to provide a fully configurable data processing pipeline with the goal of generating annotated javascript object notation (JSON) files that can be readily indexed into ElasticSearch, stored as files or pushed back to a database.

ElasticSearch

ElasticSearch[4] is an open source, broadly distributable, readily scalable enterprise search engine. It is a powerful NoSQL search engine based on the Lucene library that provides a distributed full-text search engine capable of storing large volumes of data as schema-free JSON documents. This allows it to have several advantages over structured query language (SQL). Very large complex queries in ElasticSearch, that would take over 10 seconds with SQL, will retrieve results faster by a factor of 1000 (10 milliseconds). Additionally, it performs well in searching through loosely structured, raw data enabling full-text search whereas SQL would struggle with these types of queries. However, it is important to remember that ElasticSearch is not a relational

[1] https://tika.apache.org/
[2] https://github.com/tesseract-ocr/tesseract
[3] https://spring.io/projects/spring-batch
[4] https://www.elastic.co/elasticsearch/

database and therefore single queries do not permit different table/view joins or subqueries. If these are required, this would require an annotation processing step.

The processed EPR data is stored in indices as defined in the local corresponding CogStack pipeline instance. In other words each EPR input is labelled with a unique id. Once stored, it can be rapidly queried either through the Kibana dashboard, using an ElasticSearch connector available in many programming languages, or even using one's own custom REST API.

Kibana

Kibana[5] is the Web GUI (Dashboard) of CogStack to interact with ElasticSearch. Kibana is both a data exploration tool and a data visualisation module extension for ElasticSearch that can be easily used to explore and query the data. Straight out of the box, CogStack platform deployments can use Kibana immediately as a ready-to-use data exploration tool as seen in figure 3.

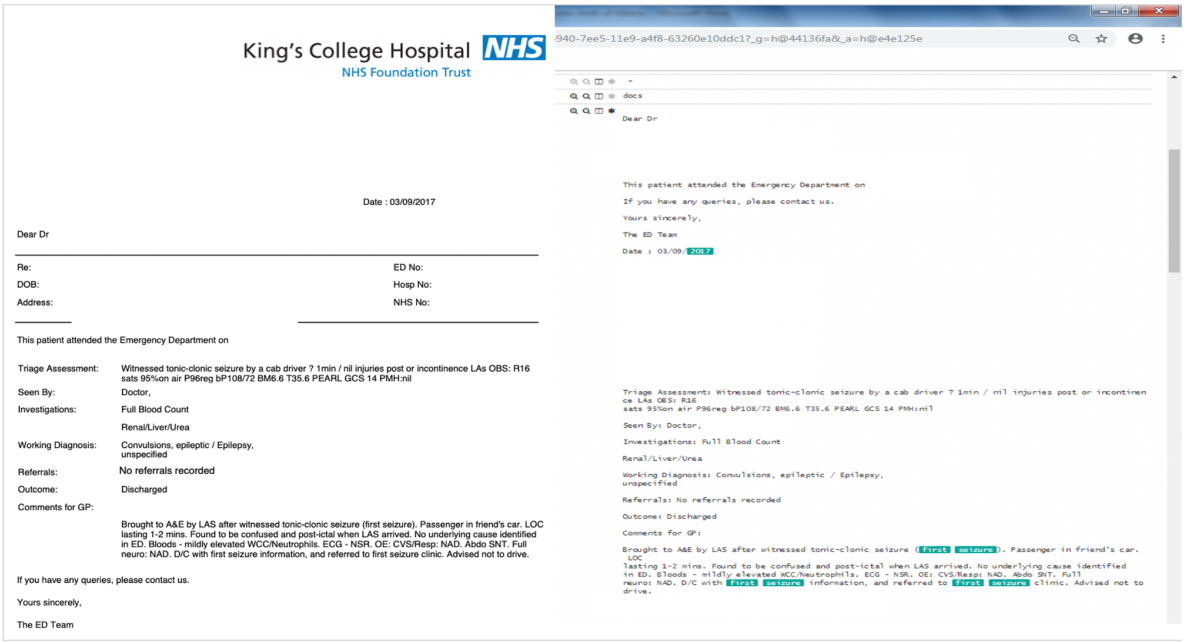


Figure 3. Display of how an EPR Document (left panel) is stripped and format standardised in CogStack stripped (right panel). Kibana dashboard highlights keyword hits in turquoise; in this example "2017" & "first seizure" to the user.

[5] https://www.elastic.co/kibana/

2.2 Python v3.6+

Python is a general-purpose, object-oriented programming language and is one of the most popular open source tools due to its high-level beginner friendly syntax (*Welcome to Python.org*, 2021). It contains many reusable chunks of code called libraries. Python has a widely developed open-source library ecosystem supported by an active community. Several popular libraries that support data science tasks were utilised, namely: NumPy[6] for handling large dimensional arrays, Pandas[7] for data manipulation and analysis, SciPy[8] for statistical analysis, RE[9] for string matching operations, and Plotly[10] for building data visualisations. This thesis also utilises a suite of specialised deep learning and machine learning libraries, including scikit-learn[11], SpaCy[12], pytorch[13], and transformers[14].

2.3 Systematized Nomenclature of MEDicine

2.3.1 What is SNOMED-CT?

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is a terminology which supports the development of comprehensive high-quality clinical content in EPR (*SNOMED CT Starter Guide*, 2021). It is currently the most comprehensive and precise clinical terminology used across the world and it is frequently used by professionals in the healthcare industry interested in the capture, retrieval, use, and analysis of clinical information. It is not just a coding system of diagnosis but covers the broad scope of clinical meanings that is relevant to record in clinical records. The terminology provides a standardised way to represent all of the different areas of clinical knowledge where each domain is represented by a 'concept' – see SNOMED-CT structure below. SNOMED-CT is a clinically validated,

[6] https://github.com/numpy/numpy
[7] https://github.com/pandas-dev/pandas
[8] https://github.com/scipy/scipy
[9] https://github.com/python/cpython/blob/3.9/Lib/re.py
[10] https://github.com/plotly/plotly.py
[11] https://github.com/scikit-learn/scikit-learn
[12] https://github.com/explosion/spaCy
[13] https://github.com/pytorch/pytorch
[14] https://github.com/huggingface/transformers

semantically rich, controlled vocabulary that is continuously maintained to capture new areas of knowledge, classification structures and to meet new emerging requirements.

SNOMED-CT international edition is actively updated with new releases every 6 months. Additionally, there are regional specific adaptations or extensions which are translated or built on top of the core international edition framework such as the SNOMED-CT UK extension. These regional editions expand upon concepts more granularly for its use and regional specific applications within the respective area of use.

2.3.2 SNOMED-CT Core Structure

There are 3 primary components which form the SNOMED terminology (*SNOMED CT Starter Guide*, 2021).

1. **Concepts** – where each represents a unique clinical meaning. All concepts within snomed are labelled with a unique numerical identifier (SCTID).
2. **Descriptions** – All concepts are also represented with a unique fully specified name (FSN) to allow them to be human readable. Alongside there are synonyms which list the alternative ways of expressing the same concept.
3. **Relationships** – There are many different relationship structures present in SNOMED. The primary one is the “is a” relationship which defines the polyhierarchical structure; this means that concepts can have one or more parent concepts and are all derived from the single SNOMED-CT root concept. This allows clinical data to be recorded at different levels of granularity and later accessed or aggregated at a higher-level group term. Other relationships include “finding by site” or mapping to ICD.

2.3.3 UK Adoption

The requirement for all UK health services to adopt SNOMED-CT is recommended by the National Information Board (NIB), in “Personalised Health and Care 2020: A Framework for Action” (*Personalised Health and Care 2020: A Framework for Action*, 2021). SNOMED-CT has already started being implemented in primary care settings since April 2018. Where all GP service provider systems must adopt SNOMED and

be used instead of READ CODES – an older, no longer supported, clinical terminology which has been used within the NHS since 1985.

Overall, SNOMED-CT is an exceptional clinical dictionary foundation to be expanded upon for clinical NLP applications. The balance between its greater breadth of coverage of clinical concepts is favourable when compared to classification systems such as the International Classification of Diseases (ICD), whilst it avoids the duplication of clinical knowledge representation found in larger terminology systems such as the Unified Medical Language system (UMLS). An example of this is 'Pervasive Development Disorder' [CUI: C0004352] and `Autistic Disorder` [CUI: C0524528].

The edition of SNOMED-CT used in this thesis was the 2020-10-01 UK extension combined with the 2020-10-01 UK drugs extension. All COVID work used the emergency interim March 2020 release that included newly added COVID-19 related concepts including any applicable changes to descriptions.

Chapter 3: First Seizure Health Trajectories and Pathways

3.1 Summary

Purpose

Suspected seizure is the most common neurological cause of admission to hospital in England. The differential diagnosis for a suspected first seizure is wide and patient management is critical to the subsequent prognosis. Electronic Patient Records (EPR) are data that are generated when patients interact with healthcare services and contain information including diagnoses, symptoms, procedures, prescriptions, and tests. This makes EPR ideal sources of information to provide evidence for the optimisation of health management. However, beyond direct patient care, its use for secondary purposes such as research or service development has so far been limited. We aimed to explore the feasibility of using EPR at King's College Hospital trust (KCH) to identify suspected first seizure patients and compare their management against guidelines as recommended by the National Institute of Clinical Excellence (NICE).

Method

We utilised a digital translational informatics platform, called CogStack, developed by King's College London to search >18 million EPR documents at KCH using rule-based natural language processing to identify suspected first seizure patients attending the emergency department in 2017. We then retrieved their EPR and extracted relevant information about their symptomatic presentation, final diagnosis, timing of investigations requested, and specialist appointments provided.

Results

226 patients attended the emergency department, of which 161 (72%) neurology outpatient appointments were provided. Common final diagnoses were epilepsy (23%), cardiovascular syncope (19%) or a single seizure or unexplained episode (19%). Our review of these records showed that in some cases the clinical pathway was not in keeping with NICE guidelines: 84% patients were not seen by a neurologist within two weeks, and slightly less than half of initial EEG and MRI investigations were not completed within four weeks as recommended by the guidelines.

Conclusion

This study has revealed that it is feasible to mine EPR at scale for the rapid analysis of service demand and the monitoring of patient health trajectories.

3.2 Authors

Anthony Shek, Javier Pena, James T. Teo, Mark P. Richardson, Eva Theochari.

Statement of Contributions

AS, JT, MPR conceived the study design
AS performed data processing and analysis
AS, JP, ET, JT performed data validation
AS, ET, JT, MPR performed critical review
AS wrote the manuscript

3.3 Introduction

Seizures are the most common neurological cause of presentation to hospitals in England with up to 100,000 admissions per year. Twenty-two percent of these are first seizure cases and therefore present a considerable burden to A&E (Dixon et al., 2015). Even having one seizure can be a traumatic physical and psychological event that can result in major adverse social consequences such as loss of driving privileges or potential limits to employment opportunities. A timely diagnosis of epilepsy is important as the occurrence of additional seizures before appropriate treatments can be initiated has been associated with poor prognosis and potential socioeconomic disadvantage (Firkin et al., 2015; Kwan & Brodie, 2000).

If a person suffers from a suspected first seizure event, they initially present either to their GP or directly to A&E. Initially, patients are often seen by paramedics, junior doctors and physicians without particular expertise in epilepsy and they often rely upon patient and bystander accounts of the events. In addition to the complexity in recognising a seizure, the differential for a suspected seizure event is broad and can be easily confused with non-seizure events related to other medical disorders e.g. psychogenic non-epileptic seizure, syncope or an unknown cause of collapse (Malmgren et al., 2012).

In the UK, the National Institute for Health and Care Excellence (NICE) guidelines offer evidence-based and expert consensus-based practice advice on managing adult patients having a first seizure (NICE, 2021a). Aligning care pathways with these guidelines offers a robust management strategy for patients and maintains consistency and quality in healthcare. The guidelines state that if after screening an epileptic seizure is suspected or if there is diagnostic doubt, the patient should be referred to a first seizure pathway, where investigations, diagnosis, and initial treatment decision are made by a specialist (who is usually, but not always, a neurologist – in some cases it might also be a general paediatrician or another specialist with particular epilepsy expertise) (NICE, 2021). The specialist would try to establish if a diagnosis of epilepsy is appropriate and then establish epilepsy type, seizure type, and probable underlying cause from a combination of clinical history, physical examination, and investigations; besides clinical history, EEGs and MRI scans are often crucial in this process (Krumholz et al., 2015). The NICE guidelines recommend that these appointments with a specialist should be considered as urgent and conducted within 2 weeks after initial presentation and that appropriate investigations occur within 4 weeks.

Each interaction between the patient and healthcare provider is recorded in the electronic patient record (EPR). These data can range from documentation of diagnoses, symptoms, procedures, prescriptions, and tests. The granularity of data makes EPR ideal sources of information to provide evidence for the optimisation of health management. However, beyond direct patient care, its use for secondary purposes such as research or service development has so far been limited. This is because electronic patient record systems are often closed, proprietary and contain incomplete or unstructured data. The result is that the wealth of information potentially available within health records is often inaccessible and underutilised. This impacts its ability to be directly used for research or service development purposes.

CogStack is an information retrieval and extraction platform developed by researchers at the NIHR Maudsley Biomedical Research Centre (BRC) (Jackson et al., 2018). CogStack implements new data mining techniques within NHS hospitals – specifically,

the ability to centralise and quickly search information held within EPR systems. Further information on CogStack can be found in this thesis' methods section.

CogStack has the potential to rapidly screen EPR to uncover how first suspected seizure patients are managed and identify patterns in their health trajectories. This will help to monitor performances of care pathways and supply the evidence for any data driven decision-making process that may be required to optimise them. However, the feasibility of using such tools is currently unknown and has not been applied to explore hospital care pathways.

3.4 Aims and Objectives

The aims of this retrospective study are (1) to explore the feasibility of using CogStack to extract clinically relevant information from the EPR system of a large NHS acute healthcare provider in England (King's College Hospital NHS Foundation Trust, KCH) to identify suspected first seizure in adult patients (>18) who presented to the emergency department in the calendar year 2017; (2) to follow their patient journey (care/clinical pathways) until appropriate investigations and specialist appointments occur; (3) to compare the timing of these appointments as well as EEG and MRI investigations to the NICE guideline standards for the management of suspected first seizure patients; and (4) to map these patients from presenting symptoms to a clinical diagnosis.

Specifically, the management of patients will be compared to the relevant quality statements in Section 1.4 of the NICE guideline standards for the management of adult first suspected seizure patients:

- **Quality statement 1**: Adults presenting with a suspected seizure are seen by a specialist in the diagnosis and management of the epilepsies within 2 weeks of presentation
- **Quality statement 2**: Adults having initial investigations for epilepsy undergo the tests within 4 weeks of them being requested.
- **Quality statement 3**: Adults who meet the criteria for neuroimaging for epilepsy have magnetic resonance imaging.

3.5 Methods

To identify suspected first seizure patients attending the KCH emergency department in 2017, we used CogStack to retrospectively search more than 18 million EPR documents at KCH using rule-based natural language processing, specifically regular expressions and fuzzy-matching techniques available through the Lucene query syntax found within the Kibana dashboard (see methods section for more information of CogStack components).

A search strategy was constructed to retrieve all A&E discharge notification documents, which contained one or more of the following keyword terms that describe a first seizure event: “First seizure”, “1st seizure”, “?seizure”, “seizure?”, “possible seizure”, “probable seizure”, “First sz”, “First siezure” (to take account of a frequent mis-spelling), “First sz”, “First fit”, “1st fit”.

The time of the A&E discharge notification was required to be uploaded to the EPR system after 1st January 2017 00:00:00 but before 1st January 2018 00:00:00. The time of A&E discharges that were stated outside of this date-time frame were excluded.

Once our cohort was established, we then retrieved all subsequent records to screen and extract information regarding their symptomatic presentation, final diagnosis, timing of EEG and MRI investigations, and timing of epilepsy specialist appointments.

The inclusion criteria for this study was:

1. Over the age of 18 at the time of the first suspected seizure episode.
2. Referred from the KCH NHS Foundation Trust emergency departments, this includes the Princess Royal University Hospital (PRUH) and King’s College Hospital, for a suspected first seizure in 2017.
3. Not already referred for investigations or to a neurologist for seizure. If many referrals were made, the date of the first referral was used. If a previous epilepsy diagnosis has been already established see 4.
4. If a previous diagnosis of epilepsy has been made, no history of seizures for at least 10 years and medication free for at least 5 years.

3.6 Results

Search Results

Figure 4 illustrates the number of documents and patients remaining after each stage of the retrieval of relevant documents, which were subsequently screened to identify our cohort. There was a total of n=226 first suspected seizure patients who met the inclusion criteria, who had an average of 15.5 records per patient, where the minimum was 1 record and the maximum were 75 records. These follow-up documents were retrieved until 1st of Jan 2019.

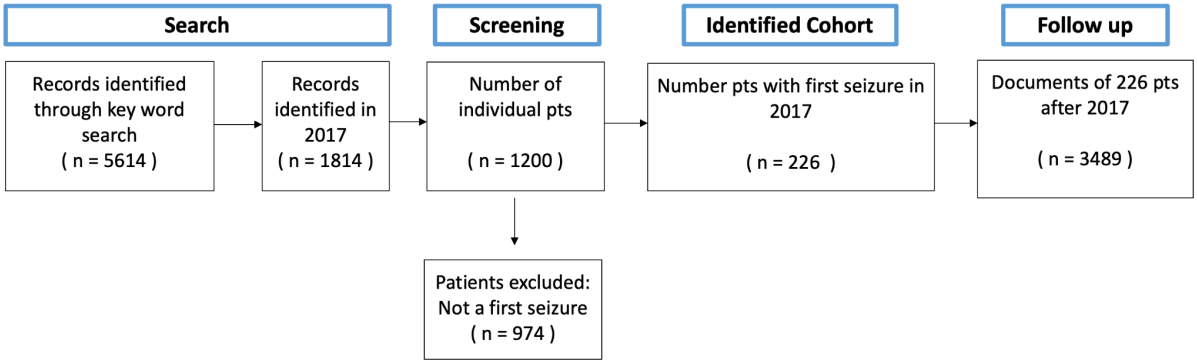


Figure 4. Flow chart of the number of documents or patients at each stage of the project.

Rate of First Suspected Seizure Referrals

In 2017, the number of referrals to an epilepsy specialist by hospital site was KCH n=166 and PRUH n=60. The monthly rate of referrals throughout 2017 was ~19 patients and the breakdown are represented in Figure 5.

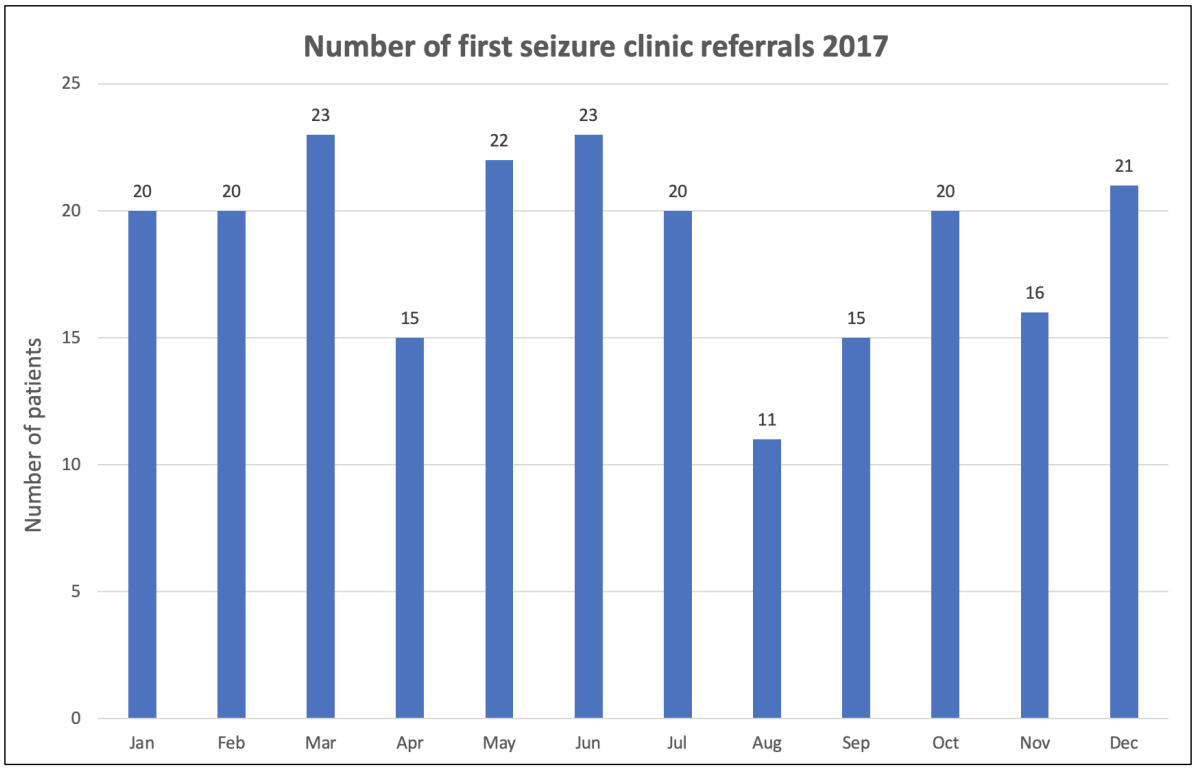


Figure 5. The frequency of 226 suspected first seizure patient referrals from KCH trust emergency departments per month within the calendar year 2017.

Patient Demographics

The demographics of these patients are represented in Figure 6. Females represented 40.3% (n=91) the mean age was 44.92 years old. Males represented a larger proportion of patients 59.7% (n=135) and were slightly younger, mean age was 43.63 years old.

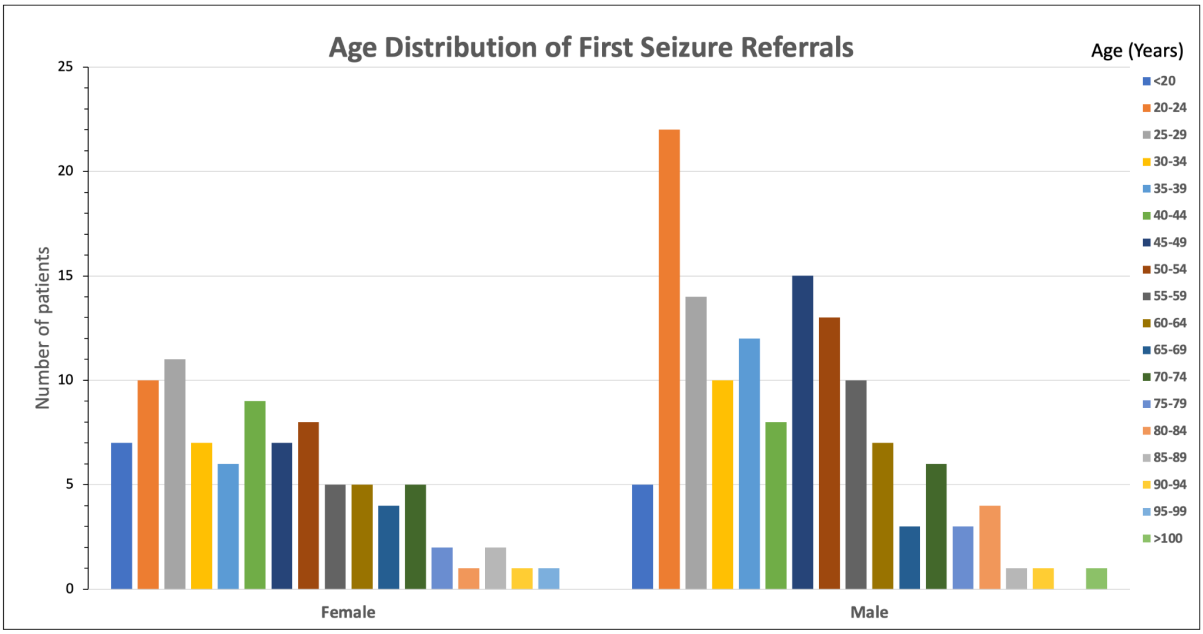


Figure 6. Gender and age distribution of n=226 first suspected seizure patients. The <20 age bracket is from 18-20 years old.

Time Until Epilepsy Specialist Appointment

From the 226 patient referrals to a neurology clinic follow-up, 161 patients were provided with a follow-up neurology appointment within KCH. Documentation reported that 12% (n=19) of these patients did not attend their appointment. Figure 7 displays that the majority of patients were seen within 6 weeks. The median time between referral and first fit clinic appointment was 9 weeks rather than 2 weeks as recommended by NICE.

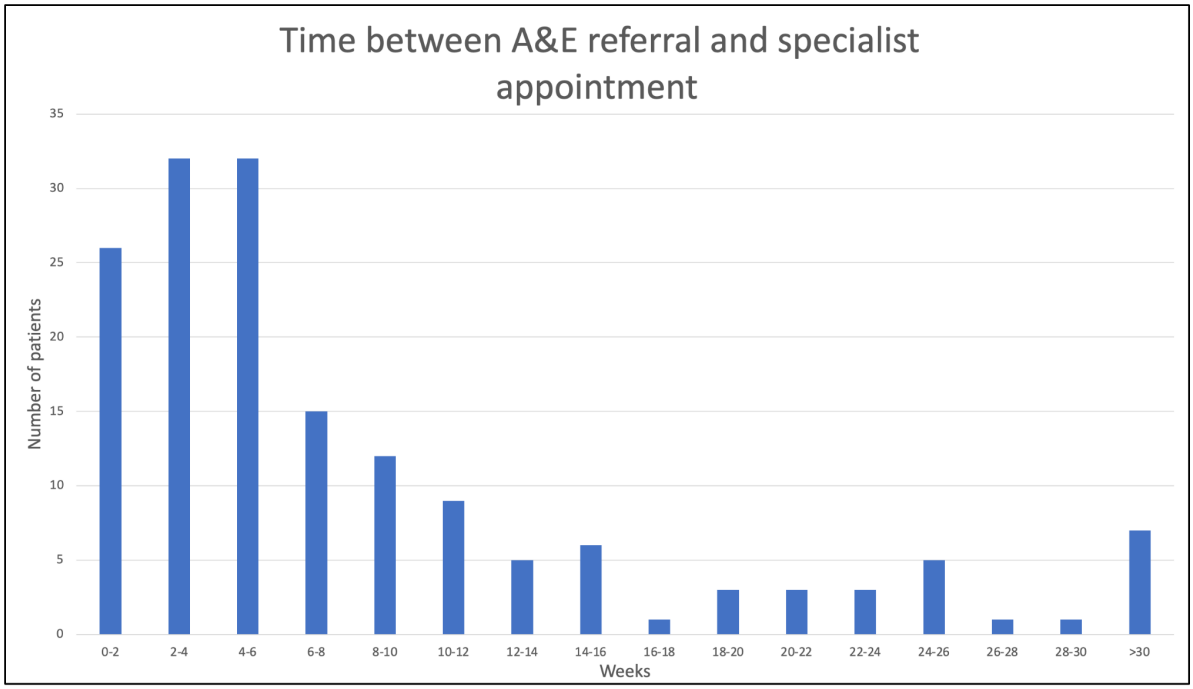


Figure 7. The time between emergency department discharge and first neurology follow-up appointment is provided in weeks.

Timing of Investigations

Only 65.8% (n=106) of patients who were referred to a neurology specialist had an EEG investigation at KCH. During emergency department admission or prior to their appointment with a neuro specialist, n=28 patients had an EEG investigation. The majority of patients had an EEG after their appointment with an epilepsy specialist n=78. The time till the EEG investigation is shown in figure 8, showing that most patients underwent an EEG investigation 5 weeks from the time of referral.

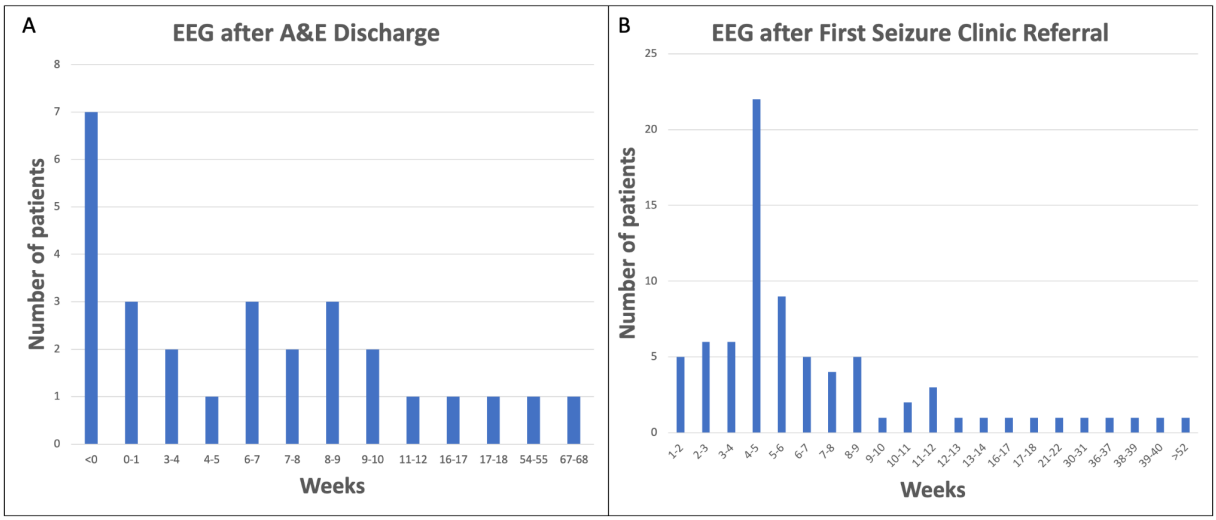


Figure 8. Time till EEG after investigation requested. Panel A represents the point of discharge from the emergency department. Panel B represents those who were referred after an initial specialist appointment. Negative number of weeks represents an EEG which occurred during the hospital admission.

The number of patients referred to have an MRI investigation was 36.4% (n=114) of patients. 36 patients had an MRI, after emergency department attendance or discharge but pre-first seizure follow-up and 78 patients had an MRI post-first follow-up (Figure 9). In both cases, the majority of patients were investigated with MRI in under 5 weeks.

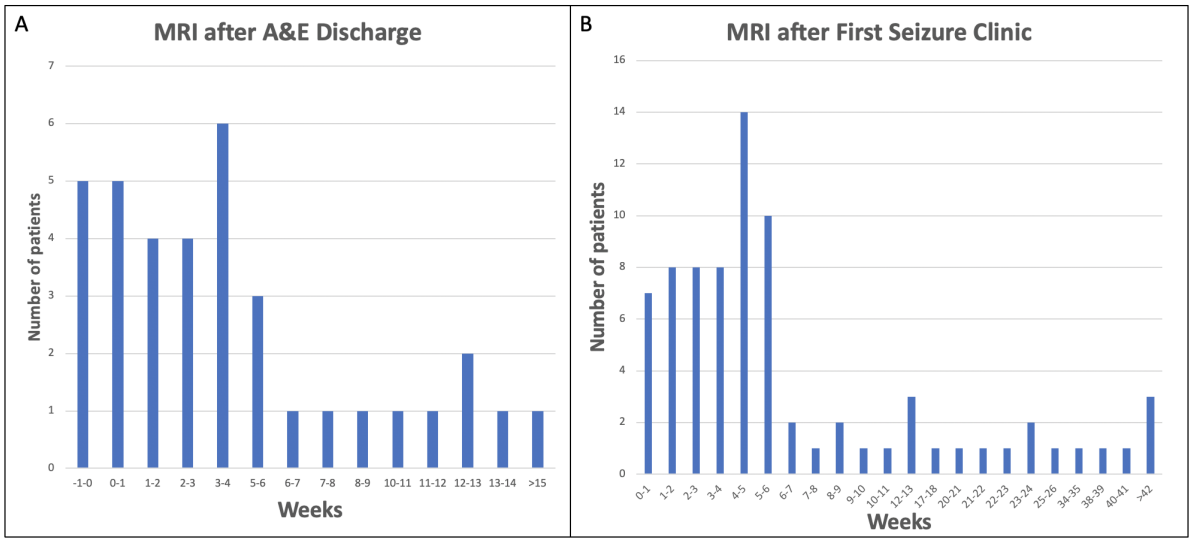


Figure 9. Time till MRI after investigation requested. Panel A represents at time from emergency department discharge. Panel B represents those who were referred after an initial specialist appointment. The negative number of weeks represents an MRI scan which took place during hospital admission.

First Seizure Trajectories

The mapping from the symptomatic reason for 161 emergency department suspected seizure patient referrals to their eventual clinical outcome are depicted in figure 10. Out of these patients referred to specialist follow-up, the major final diagnosis categories were: Epilepsy 28.0% (n=42), cardiovascular related 19.2% (n=31), or a single unknown event 18.6% (n=30). The referral based on the symptomatic presentation category ”Other” included transient global amnesia, headache, and lightheadedness. Whereas the final diagnosis category “Other” contained seizure secondary to Huntington's disease, myalgic encephalomyelitis, transient global amnesia, seizure secondary to hepatic encephalopathy. For 3.1% (n=5) of patients no diagnosis was reached a year after their first suspected seizure event and in all these cases there were ongoing discussions of whether the event was of neurological or cardiac origin. There were 12.4% (n=20) patients who attended their specialist appointment but left the hospital or were lost to follow-up before an explanation or clinical diagnosis was reached.

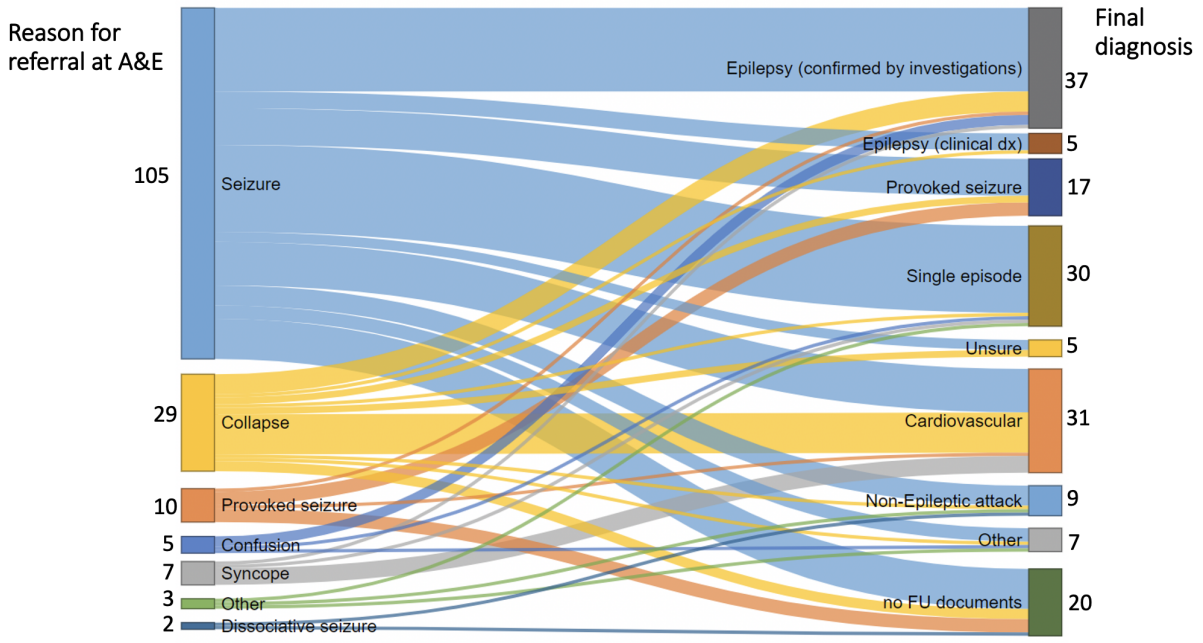


Figure 10. A Sankey plot which represents 161 first seizure patient referrals and the reason for referral from the emergency department for a suspected first seizure event (left) and their diagnostic outcome (right).

3.7 Discussion

Our findings within the KCH first suspected seizure population is in line with previous UK studies investigating similar populations of combined first seizure and epilepsy emergency department attendances (Dickson et al., 2018; Dixon et al., 2015). The 2013 national audit of seizure management in hospitals (NASH) similarly reported a larger number of male patients (57%) of average age 44 years [IQR 29–60] (Dixon et al., 2015). Whilst another national study across a 4 year period also reported slightly higher male patients attendance rate (54.6%) (Dickson et al., 2018). However, our study goes further and breaks down patient age groups shown in Figure 6 which appears to show a bimodal distribution of peak admission ages in their early 20s and late 40s. Although not explored in this study, it is possible that certain patient demographic features could be correlated with the different symptomatic reasons for referral to the first seizure pathways, and or the final diagnosis established. For example, provoked seizures where the high number of young males could be associated with alcohol.

Despite reporting similar patient demographics, previous studies were limited in study design by the fact that they used ICD-10 codes to identify their cohort using existing routinely populated databases. An example of such a database is the Hospital Episode Statistics (HES). HES is a data warehouse primarily developed for the purposes of reimbursement within the UK NHS health system. It contains routinely collected ICD-10 codes from all admissions, outpatient appointments, and emergency department attendances at NHS hospitals in England. However, the HES data quality and validity has been an area of concern with variation between under and over reporting in HES by area, hospital, and specialty (NHS Digital, 2006; Spencer & Davies, 2012; Thorn et al., 2016; Williams & Mann, 2002). Notably, HES data does not code outpatient diagnosis codes (detailed clinical data is only coded for inpatient and emergency department attendances), so HES would not be able to determine if a clinical episode of an outpatient in the first seizure clinic was diagnosed as epilepsy, cardiac syncope or some other disorder.

Additionally, ICD-10 codes do not contain the necessary granularity for the accurate extraction of suspected/possible first seizure events. NASH and Dickson et al. 2018, used epilepsy specific ICD-10 codes to identify patients with suspected seizures. Based on our study's findings, this approach would have caused an underestimation of the service demand for the first seizure care pathway as many patients who are referred to the pathway exhibit ambiguous emergency department symptomatic presentation such as “collapse” or “confusion” and they do not receive a diagnosis of epilepsy. Therefore, ICD-10 codes are often inappropriate to make robust estimates of the number of attendances related to suspected seizures.

A component of the HES is the Emergency Care Data Set (ECDS) which began data collection in 2017. This is the national data set for urgent and emergency care coding diagnoses and discharge destinations, using a limited number of SNOMED CT codes (OpenSAFELY, 2022). Future studies could use this data set to validate their own automated cohort ascertainment approach using these manually screened hospital admissions, identifying any discrepancies in code labels which can subsequently be re-reviewed.

Our study has demonstrated that CogStack is a feasible alternative to using pre-populated HES data. Enterprise search tools like CogStack have unlocked the capability to query large quantities of real-world data directly from their source rather than manually searching through an EPR system, which can be very labour intensive, or querying secondary constructed data warehouses, which have lost the granularity of the source material that it was summarising.

All patients in this study were thought to have suffered a possible first seizure event and were therefore subsequently referred to a neurological specialist for appropriate investigations and determination of the medical event (to rule out seizure or establish a diagnosis or an explanation for the event). Within the emergency department discharge notifications, a reason for the referral, based on the symptomatic presentation, was recorded. This likely reflected the assessment opinion of the emergency department clinicians during the time of referral. About a quarter of all emergency department referrals to the first seizure pathway went on to receive a diagnosis of epilepsy. However, the other large categories of diagnosis were

cardiovascular, namely cardiovascular syncopal events. This is not surprising as it is well known that these two diagnoses can easily be confused, especially when syncope is accompanied by abnormal movements such as myoclonic jerks, oral automatism, head-turning and, more rarely, urinary incontinence, thus mimicking the clinical presentation of epileptic seizures (Chowdhury et al., 2008; Ungar et al., 2017; Zaidi et al., 2000).

Comparison to NICE Guidelines

Our findings show that care at KCH frequently differed from the NICE guidelines for the management of first suspected seizure. Although NICE guidelines are evidence-based and intended to facilitate best practice, these guidelines themselves are not without their limitations. The primary data, which form the evidence for developing guidelines, are literature searches and selected expert testimonies (NICE, 2015). Susceptibility to bias relating to the nature of evidence, misconceptions, and personal recollections depending upon the beliefs of the developers, stakeholders and committee are some of the factors that may confound the real-world validity and practical deployment of these guidelines to the local context (Z. Wang et al., 2018; Woolf et al., 1999). A further difficulty arises from the generalisation that such evidence is equally applicable to every individual irrespective of variability in health urgency or circumstances (Franco et al., 2020). NICE guidelines are oriented to a single condition (e.g. epilepsy, diabetes mellitus, coronary heart disease, depression). However, patients often have several conditions at the same time and although there is guidance for multimorbidity management through NICE (NICE, 2016), this usually leads to the need to apply recommendations, potentially conflicting, from different care pathway guidelines in parallel. Therefore, these guidelines always need to be interpreted and applied using clinical judgement subject to specific settings and target groups (Z. Wang et al., 2018).

The Plan, Do, Study, Act (PDSA) cycles are a 4-stage model for service improvement, providing a framework through which any quality improvement or change to work processes can be evaluated (NHS ENGLAND, 2021). CogStack can certainly be used during the Study phase as a tool to extract information from clinical narratives to

source evidence for any future quality improvement changes to the first seizure care pathway. Control charts and proportions or funnel plots could be ways to visualise the impact of changes for each cycle.

Limitations

The main limitation of CogStack is that although string-matching (rule-based) methods are fast to write and use, they can very quickly become long and extremely complicated, even for extracting a simple idea such as a first seizure event. Additionally, it is possible that alternative ways of expressing a first seizure event, misspellings, or acronyms were not included in the search query.

This pipeline required a manual screening step to screen for a first seizure event because there were many unavoidable false positives captured in the search results. These included patients with established epilepsy revisiting emergency departments for a seizure event but were labelled as a first seizure event. Additionally, clinical documents included words and acronyms with multiple meanings (“ED”, “fit”), and frequent negation (“This is not a seizure”). This could be overcome through increasing the search and extraction pattern complexity for this use case. For example, incorporate recognition for a history of epilepsy whilst excluding negative or hypothetical mentions, and non-patient-related epilepsy mentions such as “Epilepsy clinic” or “Seen by epilepsy nurse”. Looking beyond rule-based approaches, the rapidly growing machine-learning NLP capabilities to distinguish terms based on context could potentially surpass these limitations.

The last unavoidable limitation is that any undocumented descriptions of the patient's health will not be able to be extracted. This also includes missing documents or files not uploaded to the KCH EPR system. As this study only utilised electronic records, paper records which were not digitised were not included in this study’s data pool. This affected our methodology’s ability to gather information surrounding why patients who did not receive a first seizure follow-up appointment within the trust, especially when they did not have any follow-up documents. Thereby careful document curation is an important part of the process for any second layer data analytic applications.

3.8 Conclusion

A review of the care files showed that some patients were not seen by a neurologist within 2 weeks, and initial investigations were not completed within 4 weeks, as recommended by NICE. The differential diagnosis for a suspected first seizure is wide. The most commonly reached diagnosis after a first suspected seizure event were epilepsy, cardiovascular disorders, and a single unknown episode. This work contributed to the internal redesign of the KCH first seizure pathways and acute neurology hot clinics and virtual clinics were initiated in order to improve safety, reduce variability in care, streamline investigations and management in accordance with NICE guidelines.

Overall, CogStack is a feasible tool to be used for the rapid collation of data to evaluate patient care pathways and will enable frontline clinicians caring for patients with first seizures to initiate the correct investigations and management at presentation.

Chapter 4: Multi-domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit

This chapter reproduces work which was published at Artificial Intelligence in Medicine, DOI:10.1016/j.artmed.2021.102083. It contains the methodology for a new machine-learning based approach to overcome the limitations encountered through the previous chapters use of rule-based natural language processing techniques. This chapter contains formatting adjustments which do not follow the style of the rest of the thesis.

4.1 Authors

Zeljko Kraljevic*, Thomas Searle*, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhudh, Amos A.Folarin , Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D.Shah, Wai Keong Wong, Zina Ibrahima, James T. Teo, Richard J.B. Dobson.
*These authors contributed equally.

Statement of Contributions

ZK, TS, JT, RD, AS, AF conceived the study design
ZK, TS, AS, LR, KN performed data processing and software development
ZK, TS, JT, AS, AM, LZ, ADS performed data validation
RD, JT, RS, ZI, AR, DB, ZI, RB, MPR, ADS, AM performed critical review
TS, ZK, AS, LR, ZI, RB, DB, AM, RD wrote the manuscript

Multi-domain Clinical Natural Language Processing with MedCAT: the Medical Concept Annotation Toolkit

Zeljko Kraljevic[a,1], Thomas Searle[a,f,1], Anthony Shek[c], Lukasz Roguski[b,d,h], Kawsar Noor[b,d,h], Daniel Bean[a,b], Aurelie Mascio[a,f], Leilei Zhu[d,h], Amos A. Folarin[a,d,f], Angus Roberts[a,b,f], Rebecca Bendayan[a,f], Mark P. Richardson[c], Robert Stewart[e,f], Anoop D. Shah[b,d,h], Wai Keong Wong[d,h], Zina Ibrahim[a], James T. Teo[c,g], Richard JB. Dobson[a,b,d,f,*]

[a] *Department of Biostatistics and Health Informatics Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, U.K.*
[b] *Health Data Research UK London, University College London, London, U.K.*
[c] *Department of Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, U.K.*
[d] *Institute of Health Informatics, University College London,cityLondon, U.K.*
[e] *Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London,, U.K.*
[f] *NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, U.K.*
[g] *Department of Neurology, King's College Hospital NHS Foundation Trust, London,, U.K.*
[h] *NIHR BRC Clinical Research Informatics Unit, University College London Hospitals, NHS Foundation Trust London, U.K.*

Abstract

Electronic health records (EHR) contain large volumes of unstructured text, requiring the application of Information Extraction (IE) technologies to enable clinical analysis. We present the open source Medical Concept Annotation Toolkit (MedCAT) that provides: a) a novel self-supervised machine learning algorithm for extracting concepts using any concept vocabulary including UMLS/SNOMED-CT; b) a feature-rich annotation interface for customising and training IE models; and c) integrations to the broader CogStack ecosystem for vendor-agnostic health system deployment. We show improved performance in extracting UMLS concepts from open datasets (F1:0.448-0.738 vs 0.429-0.650). Further real-world validation demonstrates SNOMED-CT extraction at 3 large London hospitals with self-supervised training over ~8.8B words from ~17M clinical records and further fine-tuning with 6K clinician annotated examples. We show strong transferability (F1>0.94) between hospitals, datasets and concept types indicating cross-domain EHR-agnostic utility for accelerated clinical and research use cases.

Keywords: Electronic Health Record Information Extraction, Clinical Natural Language Processing, Clinical Concept Embeddings, Clinical Ontology Embeddings

1. Introduction

Electronic Health Records (EHR) are large repositories of clinical and operational data that have a variety of use cases from population health, clinical decision support, risk factor stratification and clinical research. However, health record systems store large portions of clinical information in unstructured format or proprietary structured formats, resulting in data that is hard to manipulate, extract and analyse. There is a need for a platform to accurately extract information from freeform health text in a scalable manner that is agnostic to underlying health informatics architectures.

We present the Medical Concept Annotation Toolkit (MedCAT): an open-source Named Entity Recognition + Linking (NER+L) and contextualization library, an annotation tool and online learning training interface, and integration service for broader CogStack[1] ecosystem integration for easy deployment into health systems. The

*Corresponding author: richard.j.dobson@kcl.ac.uk
[1] Authors contributed equally

MedCAT library can learn to extract concepts (e.g. disease, symptoms, medications) from free-text and link them to any biomedical ontology such as SNOMED-CT[2] and UMLS[3]. MedCATtrainer[4], the annotation tool, enables clinicians to inspect, improve and customize the extracted concepts via a web interface built for training MedCAT information extraction pipelines. This work outlines the technical contributions of MedCAT and compares the effectiveness of these technologies with existing biomedical NER+L tools. We further present real clinical usage of our work in the analysis of multiple EHRs across various NHS hospital sites including running the system over 20 years of collected data pre-dating even the usage of modern EHRs at one site. MedCAT has been deployed and contributed to clinical research findings in multiple NHS trusts throughout England[5, 6].

1.1. Problem Definition

Recently NER models based on Deep Learning (DL), notably Transformers[7] and Long-Short Term Memory Networks[8] have achieved considerable improvements in accuracy[9]. However, both approaches require explicit supervised training. In the case of biomedical concept extraction, there is little publicly available labelled data due to the personal and sensitive nature of the text. Building such a corpus can be onerous and expensive due to the need for direct EHR access and domain expert annotators. In addition, medical vocabularies can contain millions of different named entities with overlaps (see Fig. 1). Extracted entities will also often require further classification to ensure they are contextually relevant; for example extracted concepts may need to be ignored if they occurred in the past or are negated. We denote this further classification as meta-annotation or a 'contextualisation' of a recognised entity. Overall, using data-intensive methods such as DL can be extremely challenging in real clinical settings.

This work is positioned to improve on current tools such as the Open Biomedical Annotator (OBA) service[10] that have been used in tools such as DeepPatient[11] and ConvAE[12] to structure and infer clinically meaningful outputs from EHRs. MedCAT allows for continual improvement of annotated concepts through a novel self-supervised machine learning algorithm, customisation of concept vocabularies, and downstream contextualisation of extracted concepts. All of which are either partially or not addressed by current tools.

1.2. NER+L in a Biomedical Context

Due to the limited availability of training data in biomedical NER+L, existing tools often employ a dictionary-based approach. This involves the usage of a vocabulary of all possible terms of interest and the associated linked concept as specified in the clinical database e.g. UMLS or SNOMED-CT. This approach allows the detection of concepts without providing manual annotations. However, it poses several challenges that occur frequently in EHR text. These include: spelling mistakes, form variability (e.g. kidney failure vs failure of kidneys), recognition and disambiguation (e.g. does 'hr' refer to the concept for 'hour' or 'heart rate' or neither).

1.3. Existing Biomedical NER+L Tools

We compare prior NER+L tools for biomedical documents that are capable of handling extremely large concept databases (completely and not a small subset). MetaMap[13] was developed to map biomedical text to the UMLS Metathesaurus. MetaMap cannot handle spelling mistakes and has limited capabilities to handle ambiguous concepts. It offers an opaque additional 'Word-Sense-Disambiguation' system that attempts to disambiguate candidate concepts that consequently slows extraction. Bio-YODIE[14] improves upon the speed of extraction compared to MetaMap and includes improved disambiguation capabilities, but requires an annotated corpus or supervised training. SemEHR[15] builds upon Bio-YODIE to somewhat address these shortcomings by applying manual rules to the output of Bio-YODIE to improve the results. Manual rules can be labour-intensive, brittle and time-consuming, but they can produce good results[16]. cTAKES[17], builds on existing open-source technologies—the Unstructured Information Management Architecture[18] framework and OpenNLP[19] the natural language processing toolkit. The core cTAKES library does not handle any of the previously mentioned challenges without additional plugins. ScispaCy[20] is a practical biomedical/scientific text processing tool, which heavily leverages the spaCy[2] library.

[2] https://github.com/explosion/spaCy

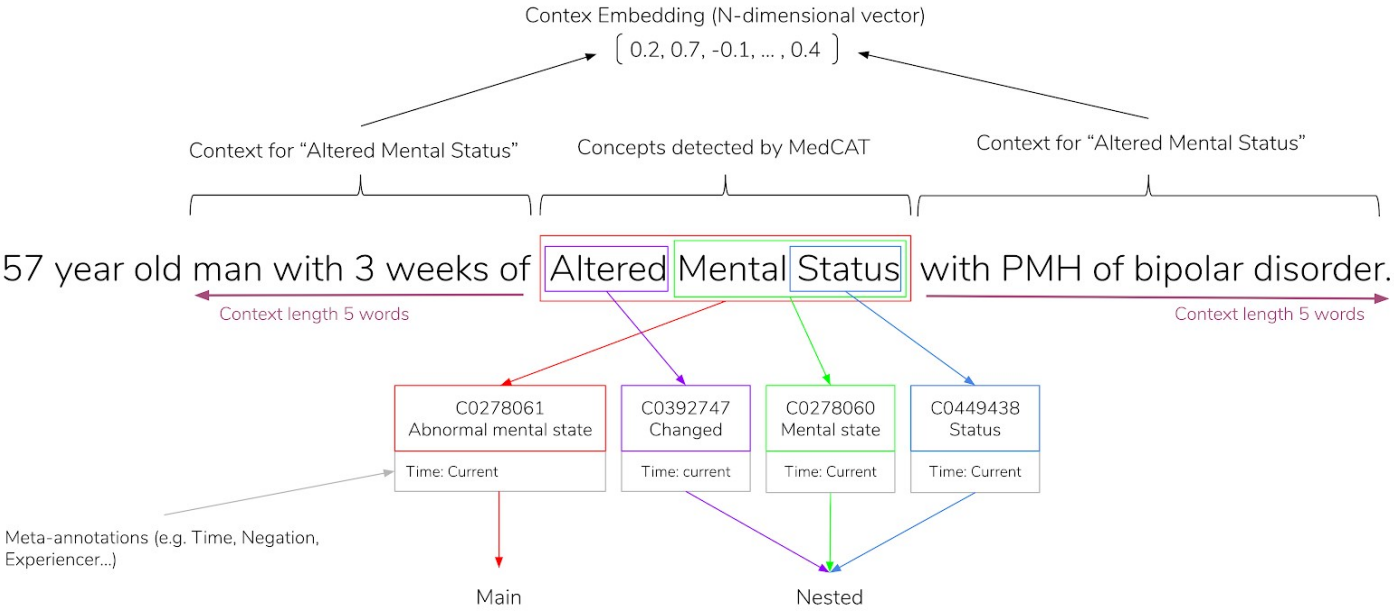


Figure 1: A fictitious example of biomedical NER+L with nested entities and further 'meta-annotations'; a further classification or 'context' applied to an already extracted concept e.g. 'time current' indicates extracted concepts are mentioned in a temporally present context. This context may also be referred to as an attribute of a recognised entity. Each one of the detected boxes (nested) has multiple candidates in the Unified Medical Language System (UMLS). The goal is to detect the entity and annotate it with the most appropriate concept ID, e.g. for the span Status, we have at least three candidates in UMLS, namely C0449438, C1444752, C1546481.

In contrast to other tools mentioned, ScispaCy is primarily a supervised model for NER with limited linking capabilities. CLAMP[21] is a comprehensive clinical NLP software that enables recognition and automatic encoding of clinical information in narrative patient reports. Similar to ScispaCy it is a supervised approach and not directly comparable to other tools mentioned here. MetaMap, BioYODIE, SemEHR, cTakes and ScispaCy only support extraction of UMLS concepts. BioPortal[22] offers a web hosted annotation API for 880 distinct ontologies. This is important for use cases that are not well supported by only the UMLS concept vocabulary[23] or are better suited to alternative terminologies[24]. However, transmitting sensitive hospital data to an externally hosted annotation web API may be prohibited under data protection legislation[25]. The BioPortal annotator is a 'fixed' algorithm so does not allow customisation or improvements through machine learning or support of non-english language corpora[26].

CLAMP, and in a limited capacity cTakes and SemEHR, support further contextualisation of extracted concepts. MetaMap, BioYODIE and scispaCy treat this as a downstream task although it is often required before extracted concepts can be used in clinical research. MedCAT addresses these shortcomings of prior tools allowing for flexibly clinician driven definition of concept contextualisation, supporting modern information extraction requirements for biomedical text.

2. Methods

MedCAT presents a set of decoupled technologies for developing IE pipelines for varied health informatics use cases. Fig. 2 shows a typical MedCAT workflow within a wider typical CogStack deployment. CogStack queries selectively extract relevant documents from the EHR including the structured and unstructured (freetext) notes. With MedCAT we firstly agree with clinical partners the relevant terms within a clinical terminology(1) and train MedCAT self-supervised(2). We load the model into the MedCATtrainer annotation tool(3) alongside a random sample of the extracted EHR documents(4). Clinical domain experts validate and improve the model using supervised online learning(5). Metrics demonstrate the quality of a fine-tuned MedCAT model(6) and once desired performance is reached the fine-tuned model is exported(7) and run upon the wider free-text EHR dataset(8,9), facilitating downstream clinical research through the newly

structured data(10).

This section presents the MedCAT platform technologies, its method for learning to extract and contextualise biomedical concepts through self-supervised and supervised learning. Integrations with the broader CogStack ecosystem are presented alongside source code[3]. Finally, we present our experimental methodology for assessing MedCAT in real clinical scenarios.

2.1. The MedCAT Core Library

We now outline the technical details of the NER+L algorithm, the self-supervised and supervised training procedures and methods for flexibly contextualising linked entities.

2.1.1. Vocabulary and Concept Database

MedCAT NER+L relies on two core components:

- **Vocabulary (VCB)**: the list of all possible words that can appear in the documents to be annotated. It is primarily used for the spell checking features of the algorithm. We have compiled our own VCB by scraping Wikipedia and enriching it with words from UMLS. Only the Wikipedia VCB is made public, but the full VCB can be built with scripts provided in the MedCAT repository (https://github.com/CogStack/MedCAT). The scripts require access to the UMLS Metathesaurus (https://www.nlm.nih.gov/research/umls).
- **Concept Database (CDB)**: a table representing a biomedical concept dictionary (e.g. UMLS, SNOMED-CT). Each new concept added to the CDB is represented by an ID and Name. A concept ID can be referred to through multiple names with identical conceptual meanings such as heart failure, myocardial failure, weak heart and cardiac failure.

2.1.2. The NER+L Algorithm

With a prepared CDB and VCB, we perform a first pass NER+L pipeline then run a trainable disambiguation algorithm. The initial NER+L pipeline starts with cleaning and spell-checking the input text. We employ a fast and lightweight spell checker (http://www.norvig.com/spell-correct.html) that uses word frequency and edit distance between misspelled and correct words to fix mistakes. We use the following rules:

- A word is spelled against the VCB, but corrected only against the CDB.
- The spelling is never corrected in the case of abbreviations.
- An increase in the word length corresponds to an increase in character correction allowance.

Next, the document is tokenized and lemmatized to ensure a broader coverage of all the different forms of a concept name. We used SciSpaCy[20], a tool tuned for these tasks in the biomedical domain. Finally, to detect entity candidates we use a dictionary-based approach with a moving expanding window:

1. Given a document d_1
2. Set window_length = 1 and word_position = 0
3. There are three possible cases:
   (a) The text in the current window is a concept in our CDB (the concept dictionary), mark it and go to 4. Note that MedCAT can ignore token order, but only for up-to two tokens (stopwords are not counted in the two token limit).
   (b) The text is a substring of a longer concept name, if so go to 4.
   (c) Otherwise reset window_length to 1, increase word_position by 1 and repeat step 3
4. Expand the window size by 1 and repeat 3.

Steps 3 and 4 help us solve the problem of overlapping entities shown in Fig. 1.

2.2. Self-Supervised Training Procedure

For concept recognition and disambiguation, we use context similarity. Initially, we find and annotate mentions of concepts that are unambiguous, (e.g. step 3. a. in the previous expanding window algorithm) then we learn the context of marked text spans. For new documents,

[3] https://cogstack.atlassian.net/wiki/spaces/COGDOC/pages/733380653/Natural+Language+Processing

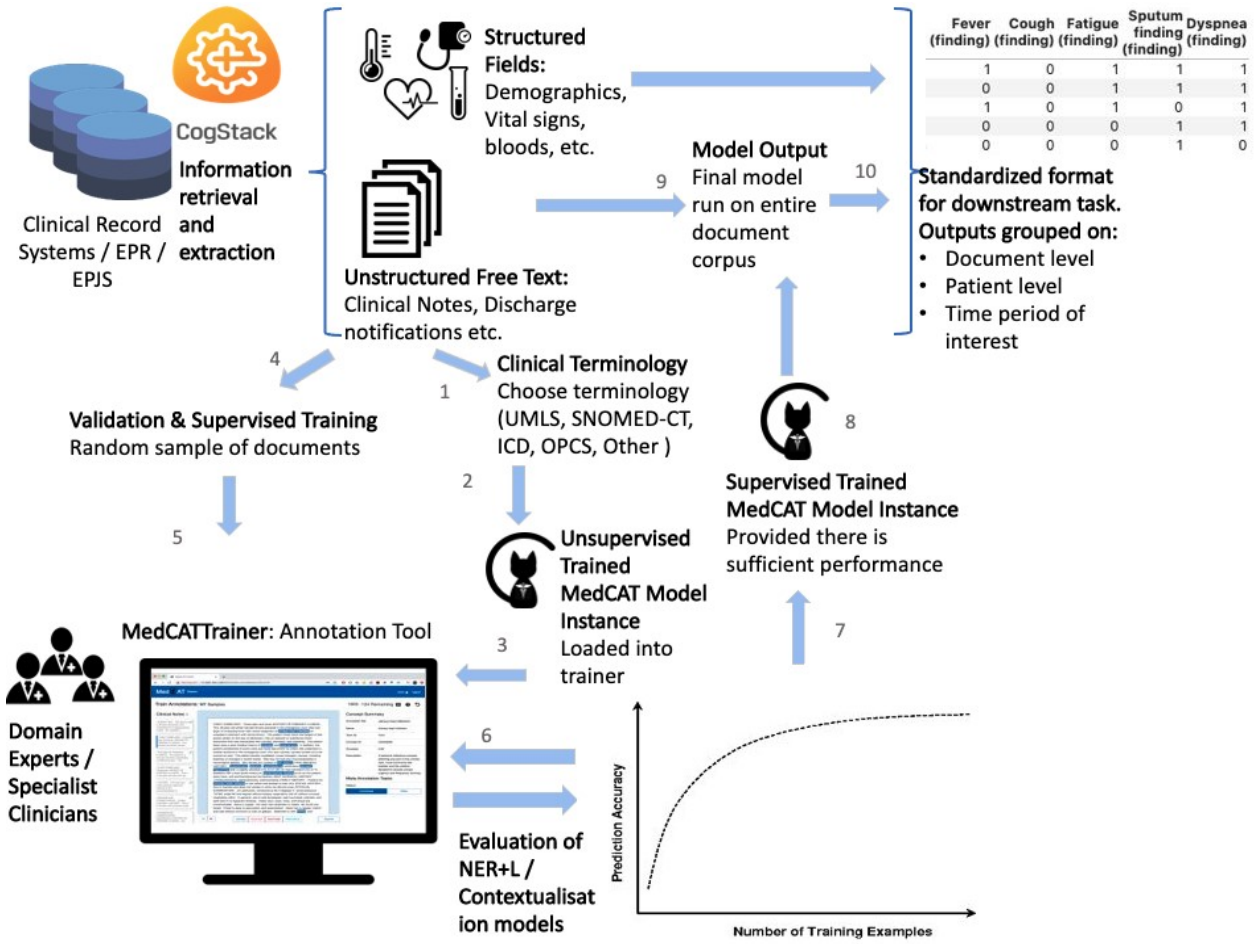


Figure 2: An example MedCAT workflow using the MedCAT core library and MedCATtrainer technologies to support clinical research.

when a concept candidate is detected and is ambiguous its context is compared to the currently learned one, if the similarity is above a threshold the candidate is annotated and linked. The similarity between the context embeddings also serves as a confidence score of the annotation and can be later used for filtering and further analysis. The self-supervised training procedure is defined as follows:

1. Given a corpus of biomedical documents and a CDB.
2. For each concept in the CDB ignore all names that are not unique (ambiguous) or that are known abbreviations.
3. Iterate over the documents and annotate all of the concepts using the approach described earlier. The filtering applied in the previous steps guarantee the entity can be annotated.
4. For each annotated entity calculate the context embedding V_{cntx}.
5. Update the concept embedding $V_{concept}$ with the context embedding V_{cntx}.

The self-supervised training relies upon one of the names assigned to each concept to be unique in the CDB. The unique name is a reference point for training to learn concept context, so when an ambiguous name appears (a name that is used for more than one concept in the CDB) it can be disambiguated. For example, the UMLS concept id:*C0024117* has the unique name Chronic Obstructive Airway Disease. This name is unique in UMLS. If we find a text span with this name we can use the surrounding text of this span for training, because it uniquely links to C0024117. ~ 95% of the concepts in UMLS have at least one unique name.

The context of a concept is represented by vector embeddings. Given a document d_1 where C_x is a detected concept candidate (Equation. 1) we calculate the context embedding. This is a vector representation of the context for that concept candidate (Equation. 2). That includes a pre-set (s) number of words to the left and right of the concept candidate words. Importantly, the concept candidate words are also included in context embedding calculation as the model is assisted by knowing what words the

surrounding context words relate to.

$$d_1 = w_1\ w_2\ \cdots\ \overbrace{w_k\ w_{k+1}}^{Cx}\ \cdots\ w_n \tag{1}$$

Where:

d_1 - An example of a document

$w_{1..n}$ - Words in the document, or to be more specific tokens

C_x - The detected concept candidate that matches the words w_k and w_{k+1}

$$V_{cntx} = \frac{1}{2s}[\sum_{i=1}^{s} V_{w_{k-i}} + \sum_{i=1}^{s} V_{w_{k+1+i}}] \tag{2}$$

Where:

V_{cntx} - Calculated context embedding

V_{w_k} - Word embedding

s - Words from left and right that are included in the context of a detected concept candidate. Typically, s is set to 9 for *long* context and 2 for *short* context.

To calculate context embeddings we use the word embedding method Word2Vec[27]. Contextualised embedding approaches such as BERT[28] were also tested alongside fastText[29] and GloVe[30]. Results presented in Section 3.1 show the BERT embeddings (the MedCAT U/MI/B configuration) perform worse on average compared to the simpler Word2Vec embeddings. FastText and GloVe perform similarly to Word2Vec, therefore our default implementation uses Word2Vec for ease of implementation. We trained 300 dimensional Word2Vec embeddings using the entire MIMIC-III[31] dataset of 53,423 admissions.

Once a correct annotation is found (a word uniquely links to a CDB name), a context embedding V_{cntx} is calculated, and the corresponding $V_{concept}$ is updated using the following formula:

$$sim = max(0, \frac{V_{concept}}{||V_{concept}||} \cdot \frac{V_{cntx}}{||V_{cntx}||}) \tag{3}$$

$$lr = \frac{1}{C_{concept}} \tag{4}$$

$$V_{concept} = V_{concept} + lr \cdot (1 - sim) \cdot V_{cntx} \tag{5}$$

Where:

$C_{concept}$ - Number of times this concept appeared during training

sim - Similarity between $V_{concept}$ and V_{cntx}

lr - Learning rate

The update rule is based on the Word2Vec model and aims to make the concept embedding $V_{concept}$ similar to the context in which the concept was presently found V_{cntx}. The scaling which is achieved via the cosine similarity is used to favour new contexts in which a concept appears over contexts that frequently appeared in the past.

To prevent the context embedding for each concept being dominated by most frequent words, we used negative sampling as defined in[27]. Whenever we update the $V_{concept}$ with V_{cntx} we also generate a negative context by randomly choosing K words from the vocabulary consisting of all words in our dataset. Here K is equal to $2s$ i.e. twice the window size for the context (s is the context size from one side of the detected concept, meaning in the positive cycle we will have s words from the left and s words from the right). The probability of choosing each word and the update function for vector embeddings is defined as:

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_j^n f(w_j)^{3/4}} \tag{6}$$

$$f(w_i) = \frac{C_{w_i}}{\sum_j^n C_{w_j}} \tag{7}$$

$$V_{ncntx} = \frac{1}{K}\sum_{i}^{K} V_{w_i} \tag{8}$$

$$sim = max(0, \frac{V_{concept}}{||V_{concept}||} \cdot \frac{V_{ncntx}}{||V_{ncntx}||}) \tag{9}$$

$$V_{concept} = V_{concept} - lr \cdot sim \cdot V_{ncntx} \tag{10}$$

Where:

n - Size of the vocabulary

$P(w_i)$ - Probability of choosing the word w_i

K - Number of randomly chosen words for the negative context

V_{ncntx} - Negative context

2.2.1. *Supervised Training Procedure*

The supervised training process is similar to the self-supervised process but given the correct concept for the

extracted term we update the $V_{concept}$ using the calculated V_{ctx} as defined in Eq. 3-10. This no longer relies upon the self-supervised constraint that at least one name in the set of possible names for a concept is unique as the correct term is provided by human annotators.

2.2.2. *Contextualisation of Identified and Linked Concepts: Meta-Annotations*

Once a span of text is recognised and linked to a concept, further contextualisation or meta-annotation is often required. For example, a simple task of identifying all patients with a fever can entail classifying the located fever text spans that are current mentions (e.g. the patient reports a fever vs the patient reported a fever but ...), are positive mentions (e.g. patient has a high fever vs patient has no sign of fever), are actual mentions (e.g. patient is feverish vs monitoring needed if fever reappears), or are experienced by the patient (e.g. pts family all had high fevers). We treat each of these contextualization tasks as distinct binary or multiclass classification tasks Meta-annotations are equivalent to 'attributes' in cTakes parlance.

The MedCAT library provides a 'MetaCAT' component that wraps a Bidirectional-Long-Short-Term-Memory (Bi-LSTM) model trainable directly from MedCATtrainer project exports. Bi-LSTM models have consistently demonstrated strong performance in biomedical text classification task[32, 33, 34] and our own recent work[35] demonstrated a Bi-LSTM based model outperforms all other assessed approaches, including Transformer models. MetaCAT models replace the specific concept of interest for example 'diabetes mellitus' with a generic parent term of the concept '[concept]'. The forward / backward pass of the model then learns a concept agnostic context representation of the concept allowing MetaCAT models to be used across concepts as observed in our results (Section. 3.3.3). The MetaCAT API follows standard neural network training methods but are abstracted away from end users whilst still maintaining enough visibility for users to understand when MetaCAT models have been trained effectively. Each training epoch displays training and test set loss and metrics such as precision, recall and F1. An open-source tutorial showcasing the MetaCAT features are available as part of the series of wider MedCAT tutorials[4]. Once trained, MetaCAT models can be exported and reused for further usage outside of initial classification tasks similarly to the MedCAT NER+L models.

2.3. *MedCATTrainer: Annotation Tool*

MedCATtrainer allows domain experts to inspect, modify and improve a configured MedCAT NER+L model. The tool either actively trains the underlying model after each reviewed document (facilitating live model improvements as feedback is provided by human users) or simply collects and validates concepts extracted by a static MedCAT model. The active learning is done on a concept level and MedCATtrainer will automatically mark some concepts as correct/incorrect and ask for user input for others where it is not confident enough. Version 0.1[4] presented a proof-of-concept annotation tool that has been rewritten and tightly integrated with the MedCAT library, whilst providing a wealth of new features supporting clinical informatics workflows. We also provide extensive documentation [5] and pre-built containers [6] updated with each new release facilitating easy setup by informatics teams.

2.4. *Datasets and Experimental Setup*

2.4.1. *Named Entity Recognition and Linking Open Datasets*

MedCAT concept recognition and linking was validated on the following publicly datasets:

1. MedMentions[36] - consists of 4,392 titles and abstracts randomly selected from papers released on PubMed in 2016 in the biomedical field, published in the English language, and with both a Title and Abstract. The text was manually annotated for UMLS concepts resulting in 352,496 mentions. We calculate that ~ 40% of concepts in MedMentions require disambiguation, suggesting a detected span of text can be linked to multiple UMLS concepts if only the span of text is considered.

[4] https://colab.research.google.com/drive/1zzV3XzFJ9ihhCJ680DaQV2QZ5XnHa06X

[5] https://github.com/CogStack/MedCATtrainer/blob/master/README.md

[6] https://hub.docker.com/r/cogstacksystems/medcat-trainer

2. ShARe/CLEF 2014 Task 2[37] - we used the development set containing 300 documents of 4 types - discharge summaries, radiology, electrocardiograms, and echocardiograms. We've used the UMLS annotations and ignored the attribute annotations.

3. MIMIC-III[31] - consists of $\sim$ 58,000 de-identified EHRs from critical care patients collected between 2001-2012. MIMIC-III includes demographic, vital sign, and laboratory test data alongside unstructured free-text notes.

We attempted to use the SemEval 2019 shared task for the evaluation of the NER+L task[7], but dataset access is currently under review for all requests to i2b2.

2.4.2. *Clinical Use Case Datasets*

Our further experiments used real world EHR data from the following UK NHS hospital Trusts:

- King's College Hospital Foundation Trust (KCH) Dataset:
  - 300 free text inpatient notes for Covid-19 positive patients, 121 Epilepsy clinic letters 2018-2019, 100 Cardiac Clinic letters, 200 echocardiographic reports, 100 CT pulmonary angiograms, 700 10k character chunks of clinical notes of patients with Diabetes Mellitus/ Gastroenteritis/ Inflammatory bowel disease/ Crohn's disease/ Ulcerative colitis for supervised training.
  - $\sim$ $17M$ documents with $\sim$ $8.8B$ tokens (entire KCH electronic health record from 1999 to 2020 consisting documents from 'multi-era', multi-vendor electronic health records (including iSoft iCM, EMIS Symphony and AllScripts) and multiple geographically-distributed hospital sites (Kings College Hospital, Princess Royal University Hospital and Orpington Hospital) were processed for self-supervised training.
- South London and Maudsley Foundation Trust (SLaM): 2200 free text notes for patients with a primary or secondary diagnosis of severe mental illness between 2007 and 2018 with each document reviewed for only a specific physical health comorbidity that may or may not appear in the note.
- University College London Hospitals Foundation Trust (UCLH) Covid-19 Datasets: 300 Free text clinical notes for Covid-19 positive or suspected patients from Jan - Apr 2020 from single-vendor electronic health record (Epic).

We used two large biomedical concept databases and prepared them as described in our source-code repository[8], the databases are:

- UMLS 2018AB: 3.82 million concepts and 14 million unique concept names from 207 source vocabularies.
- SNOMED CT UK edition: >659K concepts. The UK SNOMED CT clinical extension 20200401 and UK Drug Extension 20200325 with ICD-10 and OPCS-4 mappings.

2.4.3. *Named Entity Recognition and Linking Experimental Setup*

We use MedMentions[36], ShARe/CLEF[37] and MIMIC-III[31] datasets in our experiments. We denote the 'MedMentions' dataset (i.e. all concepts) and 'MedMentions Disorders Only' (i.e. only concepts grouped under the Disorder group as shown in[38]). We train MedCAT self-supervised on MIMIC-III configured with the UMLS database. We denote the version using Word2Vec embeddings as 'MedCAT' and the one using Bio_ClinicalBERT[39] embeddings as 'MedCAT BERT'.

An annotation by MedCAT is considered correct only if the exact text value was found and the annotation was linked to the correct concept in the CDB. We contrast our performance with the performance of tools presented in Section. 1.3. Appendix C provides self-supervised training configuration details.

2.4.4. *Clinical Use Case NER+L Experimental Setup*

For our clinical use cases we extracted SNOMED-CT terms, the official terminology across primary and sec-

[7] https://competitions.codalab.org/competitions/19350

[8] https://github.com/CogStack/MedCAT#building-concept-databases-from-scratch

ondary care for the UK National Health Health service, as this was preferred by our clinical teams over UMLS.

Fig. 3 shows our process of model training and distribution to partner hospital Trusts. Initially, we built our untrained MedCAT model using the SNOMED-CT concept vocabulary (M1), we then trained it self-supervised on the MIMIC-III dataset (M2). Next, the entire KCH EPR (17M documents with 8.8B tokens) is used for self-supervised training (M3). We collect annotations with clinician experts at KCH and train supervised (M4). We share this model with each partner hospital site where further self-supervised training (M5, M7) and specific supervised training with their respective annotation datasets (M6, M8).

Site-specific models (M3, M5, M7) are loaded into deployed instances of MedCATtrainer and configured with annotation projects to collect SNOMED-CT annotations for a range of site specific disorders, findings, symptoms, procedures and medications that our clinical teams are interested in for further research (i.e. already published work on Covid-19[5,6]). These included chronic (i.e. diabetes mellitus, ischaemic heart disease, heart failure) and acute (cerebrovascular accident, transient ischemic attack) disorders. For comparison between sites we find 14 common extracted concept groups (Appendix A) and calculate F1 scores for each concept group and reporting average, standard deviation (SD), and interquartile-range (IQR).

We shared fine-tuned MedCAT models between KCH and 2 NHS partner Trusts UCLH and SLaM. This was a collaborative effort with each hospital team only having access to their respective hospital EHR / CogStack instance. Each site collected annotated data using MedCATtrainer, tested the original base model, a self-supervised only trained model and a final supervised trained model with the MedCATtrainer collected annotations.

2.4.5. *Clinical Use Case Contextualisation Model Experimental Setup*

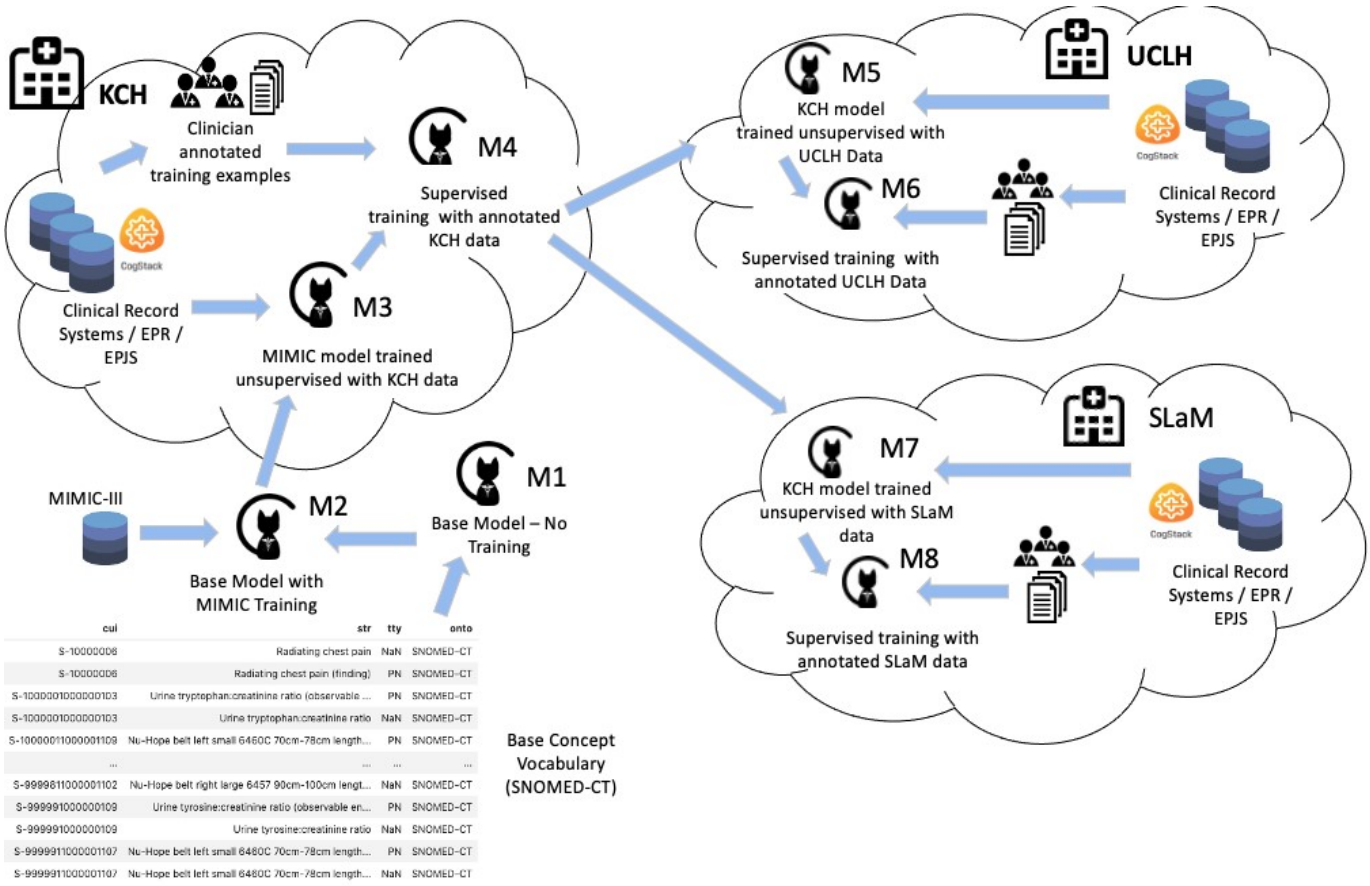
From ongoing and published work[5, 6] we configured and collected meta-annotation training examples and trained a variety of contextualisation models per site as defined in Table. 1.

Our experiments test the effectiveness of our meta annotation modelling approach to flexibly learn contextual cues by assessing cross-disorder and cross-site transferability (Section. 3.3.3. To assess cross-disorder transferability of each of the 11 disorder groups (as specified in Appendix A) we use the SLaM collected 'Diagnosis' dataset that consists of 100 annotations for each disorder group. We stratify our train/test sets by disorder, placing all examples for one disorder group in the test set and use the remaining disorder examples as a train set. We run this procedure 11 times so that each disorder group is tested once. We average all scores of each fold and report results.

Site	Task	Values
KCH	Presence	Affirmed / Negated / Hypothetical
	Experiencer	Patient / Family / Other
	Temporality	Past / Present / Future
UCLH	Negation	Yes / No
	Experiencer	Yes / No
	Problem Temporality	Past Medical Issue / Current Problem
	Certainty	Confirmed / Suspected
	Irrelevant	Yes / No
SLaM	Status	Patient / Other / NA
	Diagnosis	Yes / No

Table 1: Meta Annotation Tasks Defined Per Site, KCH = King's College Hospital NHS Foundation Trust, UCLH = University College London Hospitals NHS Foundation Trust, SLaM = South London and Maudsley NHS Foundation Trust

To demonstrate cross-site transferability we derive an equivalent meta-annotation dataset from the 'Presence' (KCH) and 'Status' (SLaM) datasets as they are semantically equivalent despite having different possible annotation values. We merge 'Presence' annotations from Affirmed/Hypothetical/False to Affirmed/Other to match classes available in SLaM. We then train and test new meta annotation models between sites and datasets report average results.



cui	str	tty	onto
S-10000006	Radiating chest pain	NaN	SNOMED-CT
S-10000006	Radiating chest pain (finding)	PN	SNOMED-CT
S-100000100000103	Urine tryptophan:creatinine ratio (observable ...	PN	SNOMED-CT
S-100000100000103	Urine tryptophan:creatinine ratio	NaN	SNOMED-CT
S-100000110000001109	Nu-Hope belt left small 6460C 70cm-78cm length...	PN	SNOMED-CT
...	...	...	...
S-9999811000001102	Nu-Hope belt right large 6457 90cm-100cm lengt...	NaN	SNOMED-CT
S-999991000000109	Urine tyrosine:creatinine ratio (observable en...	PN	SNOMED-CT
S-999991000000109	Urine tyrosine:creatinine ratio	NaN	SNOMED-CT
S-9999911000001107	Nu-Hope belt left small 6480C 70cm-78cm length...	PN	SNOMED-CT
S-9999911000001107	Nu-Hope belt left small 6480C 70cm-78cm length...	NaN	SNOMED-CT

Figure 3: Model provenance for NER+L clinical use case results between datasets and sites. M1-8, showing the MedCAT model instances, the data and method of training and base model used across all sites.

3. Results

We firstly present our concept recognition and linking results, comparing performance across previously described tools in Section. 1.3 using the UMLS concept database and openly available datasets presented in Section. 2.4 We then present a qualitative analysis of learnt concept embeddings demonstrating the captured semantics of MedCAT concepts. Finally, we show real world clinical usage of the deployed platform to extract, link and contextualise SNOMED-CT concepts across multiple NHS hospital trusts in the UK.

3.1. Entity Extraction and Linking

Table 2 presents our results for self-supervised training of MedCAT and NER+L performance compared with prior tools using openly available datasets. Metrics for all the tools were calculated consistently. Bold indicates best performance. For each manual annotation we check whether it was detected and linked to the correct Unified Medical Language System (UMLS) concept. The metrics are precision (P), recall (R) and the harmonic mean of precision and recall (F1). MedCAT models were configured with UMLS concepts and trained (self-supervised) on MIMIC-III: the base version (MedCAT) uses Word2Vec embeddings (trained on MIMIC-III), while (MedCAT BERT) uses static word embeddings from Bio_ClinicalBERT[39]. For the BERT version of MedCAT we do not use the full BERT model to calculate context representations, but only the pre-trained static word embeddings.

Our results show MedCAT improves performance compared to all prior tools across all tested metrics (excluding precision when compared to ScispaCy/CLAMP - which are supervised models). We observe that the best performance across all tools is achieved on the ShARe/CLEF dataset. However, MedCAT still improves F1 performance by 9 percentage points over the next best system. We note the simpler Word2Vec embedding (base MedCAT) on average performs better than the more expressive Bio_ClinicalBERT (BERT) embeddings. We provide a further breakdown of the range of performances by MedCAT across MedMentions and ShARe/CLEF split by UMLS semantic type in Table. 3.

Model \Dataset	**MedMentions**			**MedMentions (Disorders Only)**			**ShARe/CLEF**		
	P	R	F1	P	R	F1	P	R	F1
SemEHR	0.252	0.165	0.200	0.295	0.499	0.371	0.680	0.623	0.650
Bio-YODIE	0.316	0.143	0.197	0.445	0.366	0.402	0.700	0.607	0.650
cTAKES	0.284	0.129	0.178	0.313	0.375	0.342	0.567	0.640	0.601
MetaMap	0.305	0.465	0.368	0.358	0.460	0.403	0.755	0.540	0.630
ScispaCy*	**0.451**	0.408	0.429	0.487	0.443	0.464	0.711	0.463	0.561
CLAMP*	0.324	0.067	0.110	**0.533**	0.236	0.327	0.772	0.447	0.566
MedCAT BERT	0.386	0.475	0.426	0.459	0.513	0.485	0.788	0.678	0.729
MedCAT	0.406	**0.500**	**0.448**	0.470	**0.523**	**0.495**	**0.796**	**0.688**	**0.738**
$+\delta$ (MedCAT-Best)	-0.045	0.035	0.019	-0.063	0.024	0.031	0.041	0.048	0.088

Table 2: Comparison of NER+L tools for the extraction of UMLS concepts. *The results for ScispaCy/CLAMP are not directly comparable to other tools as they are supervised models.

Semantic Type	**Dataset**	**MedMentions**			**ShARe/CLEF**		
		P	**R**	**F1**	**P**	**R**	**F1**
T047	Disease or Syndrome	0.59	0.59	0.59	0.87	0.75	0.80
T121	Therapeutic or Preventive Procedure	0.52	0.52	0.52	NO DATA		
T061	Pharmacologic Substance	0.49	0.38	0.43	NO DATA		
T184	Sign or Symptom	0.58	0.70	0.64	0.86	0.75	0.80
T048	Mental or Behavioral Dysfunction	0.63	0.55	0.58	0.71	0.63	0.66

Table 3: MedCAT performance for different UMLS semantic types on MedMentions and ShARe/CLEF

3.2. Qualitative Analysis

For concept disambiguation the MedCAT core library learns vector embeddings from the contexts in which a concept appears. This is similar to prior work[40], although we also present a novel self-supervised training algorithm, annotation system and wider workflow. Using our learnt concept embeddings we perform a qualitative analysis by inspecting concept similarities, with the expectation that similar concepts have similar embeddings. Table. 4 shows the learnt context embeddings capture medical knowledge including relations between diseases, medications and symptoms. We train MedCAT self-supervised over MIMIC-III[31] using the entirety of UMLS, 3.82 Million concepts from 207 separate vocabularies. Training configuration details are provided in Appendix C.

3.3. Clinical Use Cases across Multiple Hospitals

The MedCAT platform was used in a number of clinical use cases providing evidence for its applicability to answer relevant, data intensive research questions. For example, we extracted relevant comorbid health conditions in individuals with severe mental illness and patients hospitalized after Covid-19 infection[5, 6, 41]. These use cases analysed data sources from 2 acute secondary/tertiary care services at King's College Hospital (KCH), University College London Hospitals (UCLH) and mental health care services South London and Maudsley (SLaM) NHS Foundation Trusts in London, UK.

The following results focus on providing an aggregate view of MedCAT performance over real NER+L clinical use-cases, meta-annotation or context classification tasks and model transferability across clinical domains (physical health vs mental health), EHR systems and concepts.

3.3.1. Entity Extraction and Linking

Table. 5 shows our results for NER+L across hospital sites, model and training configurations as described in Section 2.4.2 Our KCH annotations were collected across a range of clinicians, clinical research questions and therefore MedCATtrainer projects. This unfortunately led to a lack of resourcing to enable double annotations and calculation of inter-annotator-agreement (IIA) scores. SLaM annotations were collected by clinician / non-clinician pairs with average inter-annotator agreement (IIA) at 0.88, disagreements were discarded before results were calculated to ensure a gold-standard. UCLH IIA was at 0.85 between two medical students with annotation disagreements arbitrated by an experienced clinician providing the final gold-standard dataset. For our KCH results we use all annotations collected across various MedCATtrainer projects within our 14 concept groups as described in Section. 2.4.4 Both KCH and UCLH annotations contained occurrences of all 14 concept groups, SLaM annotated notes did not contain any occurrences of Dyspnea (SCTID:267036007), Pulmonary embolism (SCTID:59282003) and Chest pain (SCTID:29857009).

3.3.2. Entity Extraction and Linking Model Transferability

Table. 5 demonstrates the improved NER+L performance that arises from using domain specific data first self-supervised in MIMIC-III, then KCH. We observe further improvements with clinician expertise with supervised training using the KCH data. With model sharing to UCLH we observe a 0.044 average drop in F1 performance compared to KCH. Further self-supervised training directly on UCLH data offers minimal average performance gains but does reduce the F1 SD and IQR suggesting there is less variability in performance across concepts. Supervised training on a small (499) annotations from UCLH delivers comparable performance to our KCH trained model. For our experiments at SLaM we see average F1 performance drop initially by 0.062 using the KCH model directly on SLaM data. SLaM is a large mental health service provider where EHRs are markedly different to acute care hospitals KCH and UCLH. Interestingly, successive self-supervised (M7) and supervised training (M8) show benefits across all measures with final performance largely similar to final KCH performance.

Importantly, this suggests performance is transferred to the different hospital sites and initially only drops by ~0.04. With self-supervised training and further supervised training we are able to reach KCH performance with ~ 7× fewer manually collected examples at UCLH or ~ 2× fewer examples at SLaM.

3.3.3. Contextualisation Model Performance

Contextualisation of extracted and linked concepts is, by design, bespoke per project. Due to this, reporting and

Disease → Medication	Disease → Procedure	Symptom → Medication
Hypertensive disease	**Neoplastic Process**	**Fever**
Metoprolol 50 MG	Chemotherapy	Levofloxacin
Metoprolol 25 MG	Radiosurgery	Vancomycin
Valsartan 320 MG	FOLFOX Regimen	Vancomycin 750 MG
Nadolol 20 MG	Chemotherapy Regimen	Azithromycin
Atenolol 100 MG	Preoperative Therapy	Levofloxacin 750 MG
Enalapril 10 MG	Anticancer therapy	Dexamethasone
Oral form diltiazem	Parotidectomy	Lorazepam
nimodipine 30 MG	Resection of ileum	Acetaminophen

Table 4: Qualitative Analysis of Learnt Concept Embeddings. UMLS concepts that have highest cosine similarity between learnt vector embeddings of concepts in **bold**. The first row defines the chosen concept and the target concept type. We have randomly chosen the most frequent concepts and presented the 8 most similar concepts for each target concept type. For example, Neoplastic Process (C0006826) and the following rows show the top 8 most similar Procedure concepts.

Model	**Training Configuration**	**Hospital Test Site**	**# Annotated Examples**	**F1 μ**	**F1 SD±**	**F1 IQR**
M1	Base - No Training	KCH	3,358	0.638	0.297	0.333
M2	Base + Self-Supervised MIMIC-III	KCH	3,358	0.840	0.109	0.150
M3	Base + Self-Supervised KCH	KCH	3,358	0.889	0.078	0.103
M4	KCH Self-Supervised + KCH Supervised	KCH	3,358	0.947	0.044	0.051
M4	KCH Self-Supervised + KCH Supervised	UCLH	499	0.903	0.103	0.112
M5	KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised	UCLH	499	0.905	0.079	0.034
M6	KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised + UCLH Supervised	UCLH	499	0.926	0.060	0.086
M4	KCH Self-Supervised + KCH Supervised	SLaM	1,425	0.885	0.095	0.088
M7	KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised	SLaM	1,425	0.907	0.047	0.082
M8	KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised + SLaM Supervised	SLaM	1,425	0.945	0.029	0.025

Table 5: NER+L Results Across Hospitals. MedCAT NER+L performance for common disorder concepts defined in Appendix A by clinical teams. Annotations for supervised learning are used as test sets for models M1, M2, M3, M5, M7. Average performance on a 10 fold cross-validation with a held out test set is reported for models M4, M6, M8. KCH: Kings College Hospital; UCLH: University College Hospital; SLaM: South London and The Maudsley NHS Foundation Trusts.

comparing results across studies / sites is difficult as the definitions of tasks and concepts collected are different and therefore output trained models are bespoke. Table. 6a shows aggregate performance at each site, and Table. 6b-6c show further experiments for cross-site and cross-concept model transferability.

We achieve strong weighted (0.892-0.977) / macro (0.841-0.860) F1 performance across all tasks and sites, with breakdown of each metric per site/task available in Appendix D. We report average macro and weighted F1 score demonstrating the variation in performance due to unbalanced datasets across most tasks.

For cross-concept transferability, Table. 6b shows a decrease in performance when stratifying by concept. However, we still observe a relatively high 0.82-0.85 score suggesting the model is capable of learning disorder independent representations that distinguish the classification boundary for the 'Diagnosis' task, not just the disorder specific contexts.

Our cross-site transferability results, Table 6c, suggest the 'Status' context model that is trained on cross site (i.e. KCH) data then fine-tuned on site specific data (i.e. SLaM) performs better (+ 0. 0.08 Macro / + 0.09 Weighted F1) compared with training on only the SLaM site specific training only (i.e. comparing row 3 and 4).

4. Discussion

4.1. Named Entity Recognition and Linking

Our evaluation of MedCAT's NER+L method using self-supervised training was bench-marked against existing tools that are able to work with large biomedical databases and are not use-case specific. Our datasets and methods are publicly available making the experiments transparent, replicable, and extendable. With the MedMentions dataset, using only self-supervised learning, our results in 3.1, demonstrate an improvement on the prior tools for both disorder detection (F1=0.495 vs 0.464) and general concept detection (F1=0.448 vs. 0.429). We observe all tools perform best with the ShARe/CLEF dataset. We suggest this broadly due to the lack of ambiguity and the more clinical setting allowing alternative systems to also perform reasonably well.

We now discuss the result between our BERT and regular (Word2Vec) configured MedCAT models. Generally BERT, a deep neural embedding model, performs well for a range of downstream tasks[28] better than older approaches such as Word2Vec, i.e. a shallow neural embedding. We believe this due to our use of pre-trained static BERT embeddings that: 1) are not specifically trained to produce similar values for words appearing in a similar context, 2) Sub-word tokenization might be problematic if the tokenizer was trained on a non-medical dataset (no matter whether it was fine-tuned later on MIMIC-III, pubmed or similar).

The general concept detection task with MedMentions is difficult due to: the larger number of entities to be extracted, the rarity of certain concepts and the often highly context dependent nature of some occurrences. Recent work[42] highlights examples of ambiguous texts within the MedMentions dataset such as 'probe' with 7 possible labels ('medical device', 'indicator reagent or diagnostic aid' etc.) Further work[40] also showed a deep learning approach (BioBERT+) that achieved F1=0.56. When MedCAT is provided with the same supervised training data we achieve F1=0.71. We find our improved performance is due to the long tail of entities in MedMentions that lack sufficient training data for methods such BioBERT to perform well.

Our qualitative inspection of the learnt concept embeddings, 3.2 indicate learnt semantics of the target medical domain. This result mirrors similar findings reported in fields such as materials science[43]. Recent work has suggested an approach to quantity the effectiveness of learnt embeddings[38] in representing the source ontology. However, this relies on concept relationships to be curated before assessment requiring clinical guidance that may be subjective in the clinical domain. We leave a full quantitative assessment of the learnt embeddings to future work for this reason.

As more concepts are extracted the likelihood of concepts requiring disambiguation increases, particularly in biomedical text[44]. Estimating the number of training samples for successful disambiguation is difficult but based on our experiments we need at least 30 occurrences of a concept in the free text to perform disambiguation. We provide more details in Appendix B.

Finally, we note that there are no limitations algorithmically for MedCAT to support languages other than our tested language, English. As MedCAT uses a concept dictionary/vocab for NER+L, if there are existing re-

Table 6: Contextualisation Model Results

Site	Task	# Annotated examples	Macro F1	Weighted F1
KCH	Presence	37,310	0.846	0.929
	Temporality	18,670	0.803	0.943
	Experiencer	18,670	0.867	0.959
SLaM	Patient Diagnosis	1,152	0.904	0.913
	Status	1,152	0.775	0.812
UCLH	Negation	4,400	0.836	0.970
	Experiencer	4,400	0.940	0.996
	Problem Temporality	4,350	0.848	0.970
	Certainty	4,160	0.836	0.970
	Irrelevant	4,390	0.835	0.969

(a) Site Specific Contextualisation Model Performance. Weighted / Macro average F1 Meta annotation model performance custom defined and trained per site - detailed definitions are provided in Appendix D. Task definitions are uniquely defined at each site, e.g. Experiencer at KCH considers the values patient / family / other whereas Experiencer at UCLH only considers the value patient / other. Status at SLaM considers the values affirmed / other and Certainty at UCLH considers the values confirmed / suspected. We include all concepts of interest as defined under clinician guidance at each site, therefore site-to-site comparison in performance cannot be made.

Site	Task	Train / Test Split	Macro F1	Weighted F1
SLaM	Diagnosis	Concept Stratified	0.82	0.85
SLaM	Diagnosis	Random	0.90	0.91

(b) Cross Site Transferability Performance. 11 fold concept stratified CV vs randomized CV for SLaM 'Diagnosis' contextualisation task performance. The 11 concepts were selected from NER+L experiment concepts available at SLaM (Supplementary Table 1). The 'Diagnosis' task at SLaM was used as this was our most balanced dataset between all tasks and concepts collected.

Site	Trained on	# Annotated Examples	Macro F1	Weighted F1
KCH	KCH	37,310	0.89	0.93
SLaM	KCH	37,310	0.71	0.91
SLaM	SLaM	1,152	0.77	0.87
SLaM	KCH + SLaM	38,462	0.85	0.96

(c) Cross-site transferability of the MetaCAT model for Presence (at KCH) / Status (at SLaM converted to values of Affirmed/Other) - as that was the only task that existed across sites. Results show 10 fold CV where applicable - e.g. row 2 is direct testing of the KCH model on SLaM data, so no training is performed on the SLaM side.

sources (e.g. SNOMED-CT has already been translated into Spanish, Dutch, Swedish and Danish) they can be used directly for these languages with likely similar results. Alternatively, users could build their own custom concept dictionary (CDB) for their language of choice. Meta-annotation or contextualisation models also do not have language specific features, i.e. English, and would also likely perform well as they only rely on bi-directional context from supervised examples to make predictions.

4.2. Clinical Use Cases

MedCAT models and annotated training data have been implemented to be easily shared and reused, facilitating a federated learning approach to model improvement and specialisation with models brought to sensitive data silos. Our results in Section. 3.3 demonstrate that we are able directly apply models trained at one hospital site (KCH) to multiple other sites, and clinical domains (physical vs mental health datasets) with only a small drop in average F1 (0.044 at UCLH, 0.062 at SLaM), and after small amount of additional site specific training, we observe comparable performance (-0.021 at UCLH, -0.002 at SLaM).

We also highlight that separate teams were able to deploy, extract and analyse real clinical data using the tools as is by following provided examples, documentation and integrations with the wider CogStack ecosystem. Academic engineering projects are often built to support a single research project, however MedCAT and the CogStack ecosystem are scalable fit-for-purpose locally-tunable solutions for teams to derive value from their data instead of being stalled by poor quality code or lack of documentation. This means the model is broadly useful with top-up tuning also available for specific scenarios, domains and hospitals.

Each hospital site and clinical team freely defined the set of contextualisation tasks and associated values for each task. On aggregate our results show performance is consistently strong across all sites and tasks (Macro F1: 0.841-0.860, Weighted F1: 0.892-0.977). With many of the tasks the annotated datasets are highly unbalanced. For example, the 'Presence' task at KCH, disorders are often only mentioned in the EHR if they are affirmed (e.g. "...pmhx: TIA..."), and only rarely are hypothetical (e.g. "...patient had possible TIA...") or negated terms (e.g. "...no sign of TIA...") encountered. This explains the differences in performance when reporting macro vs weighted average F1 score. We would expect generalization performance to lie between these reported metrics.

4.3. Limitations

MedCAT is able to employ a self-supervised training method as the initial pass of the algorithm uses a given unique name to learn and improve an initial concept embedding. However, if the input vocabulary linked to the concepts inadequately specifies possible names or the given names of a concept rarely appear in the text then improvements can only occur during standard supervised learning. The main limitation of our approach is that it greatly depends on the quality of the concept database. Large biomedical concept databases (e.g. UMLS) however have a well specified vocabulary offering many synonyms, acronyms and differing forms of a given concept.

A limitation of our concept embedding approach is if different concepts appear in similar contexts disambiguation and linking to the correct concept can be difficult. For example, 'OD' can link to 'overdose' or 'once daily', both referring to medications with very different implications. We have rarely seen this problem during real-world corpus. Our approach can also struggle if concepts appear in many varying contexts that are rarely seen or annotated for. With each new context updating the underlying concept embedding this may decrease performance of the embedding.

Supervised learning requires training data to be consistently labelled. This is a problem in the clinical domain that consists of specialised language that can be open to interpretation. We recommend using detailed annotation guidelines that enumerate ambiguous scenarios for annotators.

4.4. Future Work

MedCAT uses a vocabulary based approach to detect entity candidates. Future work could investigate the expansion of such an approach with a supervised learning model like BERT[28]. The supervised learning model would then be used for detection of entity candidates that have enough training data and to overcome the challenge of detecting new unseen forms of concept names. The vocabulary based approach would cover cases with insufficient annotated training data or concepts that have few

different names (forms). The linking process for both approaches would remain the same self-supervised.

Our self-supervised training over the ~20 year KCH EHR, as described in Section. 2.4, took over two weeks to complete. Future work could improve the training speed by parallelizing this process since concepts in a CDB are mostly independent of one another. Further work could address effective model sharing, allowing subsequent users/sites to benefit from prior work, where only model validation and fine-tuning is required instead of training from scratch.

Finally, ongoing work aims to extend the MedCAT library to address relation identification and extraction. For example, linking the extracted drug dosage / frequency with the associated drug concept, or identifying relations between administered procedures and following clinical events.

5. Conclusions

This paper presents MedCAT a multi-domain clinical natural language processing toolkit within a wider ecosystem of open-source technologies namely CogStack.

The biomedical community is unique in that considerable efforts have produced comprehensive concept databases such as UMLS and SNOMED-CT amongst many others. MedCAT flexibly leverages these efforts in the extraction of relevant data from a corpus of biomedical documents (e.g. EHRs). Each concept can have one or more equivalent names, such as abbreviations or synonyms. Many of these names are ambiguous between concepts. The MedCAT library is based upon a simple idea: at least one of the names for each concept is unique and given a large enough corpus that name will be used in a number of contexts. As the context is learned from the unique name, when an ambiguous name is later detected, its context is compared to the learnt context, allowing us to find the correct concept to link. By comparing the context similarity we can also calculate confidence scores for a provided linked concept.

With MedCAT we have built an effective, high performance IE algorithm demonstrating improved performance over prior solutions on open access datasets. We have commoditised the development, deployment and implementation of IE pipelines with supporting technologies MedCATtrainer / MedCATservice supporting the transfer, validation, re-use and fine-tuning of MedCAT models across sites, clinical domains and concept vocabularies. MedCAT deployments are enabled by extensive documentation, examples, APIs and supporting real world clinical use cases outlined in prior published work.

Overall, MedCAT is built to enable clinical research and potential improvements of care delivery by leveraging data in existing clinical text. Currently, MedCAT is deployed in a number of hospitals in the UK in silo or as part of the wider CogStack ecosystem, with wide-ranging use cases to inform clinical decisions with real-time alerting, patient stratification, clinical trial recruitment and clinical coding. The large volume of medical information that is captured solely in free text is now accessible using state-of-the-art healthcare specific NLP.

Acknowledgements

We would like to thank all the clinicians who provided annotation training for MedAT; this includes Rosita Zakeri, Kevin O'Gallagher, Rosemary Barker, David Nicholson Thomas, Rhian Raftopoulos, Pedro Viana, Elisa Bruno, Eugenio Abela, Mark Richardson, Naoko Skiada, Luwaiza Mirza, Natalia Chance, Jaya Chaturvedi, Tao Wang, Matt Solomon, Charlotte Ramsey and James Teo.

RD's work is supported by 1.National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. 2. Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. 3. The National Institute for Health Research University College London Hospitals Biomedical Research Centre.

DMB is funded by a UKRI Innovation Fellowship as part of Health Data Research UK MR/S00310X/1 (https://www.hdruk.ac.uk).

RB is funded in part by grant MR/R016372/1 for the King's College London MRC Skills Development Fel-

lowship programme funded by the UK Medical Research Council (MRC, https://mrc.ukri.org) and by grant IS-BRC-1215-20018 for the National Institute for Health Research (NIHR, https://www.nihr.ac.uk) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London.

ADS is supported by a postdoctoral fellowship from THIS Institute. AS is supported by a King's Medical Research Trust studentship. RS is part-funded by: i) the National Institute for Health Research (NIHR) Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London; ii) a Medical Research Council (MRC) Mental Health Data Pathfinder Award to King's College London; iii) an NIHR Senior Investigator Award; iv) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.

This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust, The UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare (AI4VBH); the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, MRC, NIHR or the Department of Health and Social Care. We thank the patient experts of the KERRI committee, Professor Irene Higginson, Professor Alastair Baker, Professor Jules Wendon, Professor Ajay Shah, Dan Persson and Damian Lewsley for their support.

Data Availability

Data for reproduction of experiments for the assessment for the core NER+L in comparison with are available from prior work (MedMentions, ShARe/CLEF 2014 Task 2, MIMIC-III). Due to the confidential nature of free-text data, we are unable to make patient-level data available. Interested readers should contact the authors to discuss feasibility of access of de-identified aggregate data consistent with legal permissions.

Code Availability

All code for running the experiments, the toolkit and integration with wider CogStack deployments are available here:

MedCAT: `https://github.com/CogStack/MedCAT`

MedCAT Tutorials/Example Code: `https://github.com/CogStack/MedCAT/tree/master/tutorial`

MedCATtrainer: `https://github.com/CogStack/MedCATtrainer`

MedCATtrainer Examples: `https://github.com/CogStack/MedCATtrainer/tree/master/docs`

MedCATservice: `https://github.com/CogStack/MedCATservice`

CogStack: `https://github.com/CogStack/CogStack-Pipeline`

Declaration of Interests

JTHT received research support and funding from InnovateUK, Bristol-Myers-Squibb, iRhythm Technologies, and holds shares ¡£5,000 in Glaxo Smithkline and Biogen.

Data Access Ethics

NER+L experiments use freely available open-access datasets accessible by data owners. SNOMED-CT and UMLS licences were obtained by all users at all hospital sites. Site specific ethics is listed below. KCH: This project operated under London South East Research Ethics Committee approval (reference 18/LO/2048) granted to the King's Electronic Records Research Interface (KERRI); specific work on research on natural language processing for clinical coding was reviewed with expert patient input on the KERRI committee with Caldicott Guardian oversight. Direct access to patient-level data is not possible due to risk of re-identification, but aggregated de-identified data may be available subject to legal permissions. UCLH: UCLH is deploying CogStack within its records management infrastructure and is growing its capacity to annotate its clinical records as part of wider work for routine curation. The work at UCLH described here is a service evaluation

that represents MedCAT's annotation of the records. Access to the medical records will not be possible given their confidential nature. SLaM: This project was approved by the CRIS Oversight Committee which is responsible for ensuring all research applications comply with ethical and legal guidelines. The CRIS system enables access to anonymised electronic patient records for secondary analysis from SLaM and has full ethical approvals. CRIS was developed with extensive involvement from service users and adheres to strict governance frameworks managed by service users. It has passed a robust ethics approval process acutely attentive to the use of patient data. Specifically, this system was approved as a dataset for secondary data analysis on this basis by Oxfordshire Research Ethics Committee C (08/H06060/71). The data is de-identified and used in a data-secure format and all patients have the choice to opt-out of their anonymized data being used. Approval for data access can only be provided from the CRIS Oversight Committee at SLaM.

Author Contributions

ZK, TS, JT, RD, AS, AF conceived the study design

ZK, TS, AS, LR, KN performed data processing and software development

ZK, TS, JT, AS, AM, LZ, ADS performed data validation

RD, JT, RS, ZI, AR, DB, ZI, RB, MPR, ADS, AM performed critical review

TS, ZK, AS, LR, ZI, RB, DB, AM, RD wrote the manuscript

Appendices

Appendix A. SNOMED-CT Groupings

Each group was defined with expert clinical guidance. S-267036007 - Dyspnea (finding), S-59282003 - Pulmonary embolism, (disorder) S-29857009 - Chest pain (finding) do not appear in the SLaM annotations for supervised training.

Table A.1: SNOMED-CT concept level groupings for clinical use cases

Container Concept	Concepts
S-73211009 - Diabetes mellitus(disorder)	S-44054006 - Diabetes mellitus type 2 (disorder) S-46635009 - Diabetes mellitus type 1 (disorder) S-422088007 - Disorder of nervous system co-occurrent and due to diabetes mellitus (disorder) S-25093002 - Disorder of eye co-occurrent and due to diabetes mellitus (disorder) S-73211009 - Diabetes mellitus (disorder)
S-84114007 -Heart failure (disorder)	S-128404006 - Right heart failure (disorder) S-48447003 - Chronic heart failure (disorder) S-56675007 - Acute heart failure (disorder) S-85232009 - Left heart failure (disorder) S-42343007 - Congestive heart failure (disorder) S-84114007 - Heart failure (disorder)
S-414545008 - Ischemic heart disease (disorder)	S-413439005 - Acute ischemic heart disease (disorder) S-413838009 - Chronic ischemic heart disease (disorder) S-194828000 - Angina (disorder) S-22298006 - Myocardial infarction (disorder) S-414545008 - Ischemic heart disease (disorder)
S-38341003 - Hypertensive disorder, systemic arterial (disorder)	S-31992008 - Secondary hypertension (disorder) S-48146000 - Diastolic hypertension (disorder) S-56218007 - Systolic hypertension (disorder) S-59621000 - Essential hypertension (disorder) S-38341003 - Hypertensive disorder systemic arterial (disorder)
S-13645005 - Chronic obstructive lung disease (disorder)	S-195951007 - Acute exacerbation of chronic obstructive airways disease (disorder) S-87433001 - Pulmonary emphysema (disorder) S-13645005 - Chronic obstructive lung disease (disorder)
S-195967001 - Asthma (disorder)	S-195967001 - Asthma (disorder)
S-709044004 - Chronic kidney disease (disorder)	S-723190009 - Chronic renal insufficiency (disorder) S-709044004 - Chronic kidney disease (disorder)
S-230690007 - Cerebrovascular accident (disorder)	S-25133001 - Completed stroke (disorder) S-371040005 - Thrombotic stroke (disorder) S-371041009 - Embolic stroke (disorder) S-413102000 - Infarction of basal ganglia (disorder) S-422504002 - Ischemic stroke (disorder) S-723082006 - Silent cerebral infarct (disorder) S-1078001000000105 - Haemorrhagic stroke (disorder) S-230690007 - Cerebrovascular accident (disorder)
S-266257000 - Transient ischemic attack (disorder)	S-266257000 - Transient ischemic attack (disorder)
S-84757009 - Epilepsy (disorder)	S-352818000 - Tonic-clonic epilepsy (disorder)

Continued on next page

Table continued from previous page

Container Concept	Concepts
	S-19598007 - Generalized epilepsy (disorder) S-230456007 - Status epilepticus (disorder) S-509341000000107 - Petit-mal epilepsy (disorder) S-84757009 - Epilepsy (disorder)
S-49436004 - Atrial fibrillation (disorder)	S-49436004 - Atrial fibrillation (disorder)
S-267036007 - Dyspnea (finding)	S-267036007 - Dyspnea (finding)
S-59282003 - Pulmonary embolism (disorder)	S-59282003 - Pulmonary embolism (disorder)
S-29857009 - Chest pain (finding)	S-29857009 - Chest pain (finding)

Number of examples per concept	F1 on Test
1	0.74
5	0.81
10	0.82
30	0.86

Table B.2: Relation between the number of training examples and performance of MedCAT concept disambiguation.

Appendix B. Estimating Example Counts for Sufficient F1 Score

To test the required number of examples to achieve a high enough F1 score, we created a mini-dataset from MedMentions. It contains two concepts: C0018810 (Heart Rate) and C2985465 (Hazard Ratio). Both concepts have a unique name and the ambiguous abbreviation HR that can link to either one. We chose these two concepts, as the abbreviation HR is the most frequent ambiguous concept in MedMentions, given the requirement that it must be ambiguous. Our dataset consists of:

- 60 training examples (30 per concept). In each example the full name of the concept was used, see below MedMentions Text Extracts.
- 174 test examples, each document contains the ambiguous abbreviation HR, see below MedMentions Text Extracts.

We have tested the performance for different sizes of the training set: 1, 5, 10 and 30. If we set the training set size to e.g. 5, we split the full training set into 6 parts (in total the training set has 30 examples per concept), each containing 5 examples per concept. Then we check the performance for each part and report the average over the 6 parts, see Table B.2.

Appendix C. Self-Supervised Training Configuration

Appendix C.1. Self-Supervised Training Configuration

MedCAT was configured for self-supervised training across experiments presented in Section. 2.1 as follows:

- Misspelled words were fixed only when 1 change away from the correct word for words under 6 characters, and 2 changes away for words above 6 characters.
- For each concept we calculate long and short embeddings and take the average of both. The long embedding takes into account s = 9 words from left and right (as shown in Equation 2). The short embedding takes into account s = 2 words from left and right. The exact numbers for s were calculated by testing the performance of all possible combinations for s in the range [0, 10].
- The context similarity threshold used for recognition is 0.3 unless otherwise specified. This means for a given concept candidate, or sequence of words, to be recognised and linked to the given concept the concept similarity provided by Equation 2 would be greater than 0.3.

Appendix C.2. Qualitative Analysis Training Configuration

We train MedCAT self-supervised over MIMIC-III using the entirety of UMLS, 3.82 Million concepts from 207 separate vocabularies. We use 2.4M clinical notes (nursing notes, notes by clinicians, discharge reports etc.) on a small one-core server taking approximately 30 hours to complete.

Appendix D. Contextualisation Task Results Per Site

Appendix D.1. Contextualisation Results Breakdown for KCH

Aggregate results for each defined meta-annotation at KCH. Performance is aggregated over all extracted concepts listed in Appendix A. We defined the following meta-annotation tasks:

- Presence: is the concept affirmed, negated or hypothetical, values: [Affirmed, Negated, Hypothetical]
- Experiencer: is the concept experienced by the patient or other, values: [Patient / Family / Other]
- Temporality: is the concept in the past, present or future, values: [Past, Recent, Future]

Levels of fibrin degradation products (FDP), D-dimer, fibrinogen, the ratio of FDP to fibrinogen, the ratio of D-dimer to fibrinogen, systolic blood pressure, **heart rate**, the Glasgow Coma Scale, pH, base excess, hemoglobin and lactate levels, the pattern of pelvic injury, and injury severity score were measured at hospital admission, and compared between the two groups.

NEAC was assessed by a validated food frequency questionnaire collected at baseline. We categorized the distribution of NEAC into sex - specific quartiles and used multivariable adjusted Cox proportional hazards regression models to estimate **hazard ratios** with 95% confidence intervals (95% CI).

In the overall population radical nephrectomy was not associated with an increased risk of other cause mortality on multivariable analysis compared to nephron sparing surgery (**HR** 0.91, 95% CI 0.6-1.38, p = 0.6).

Figure B.4: MedMentions Text Extracts: Three samples from the dataset used to test the amount of training samples needed for disambiguation to work. First example is a training case for the concept C0018810, second for C2985465 and third is used to test the disambiguation performance.

Table D.3: Meta Annotation Results at KCH

CLS	F	P	R	Support Test (10% of total)
Hypothetical	0.756	0.797	0.72	360
Negated	0.865	0.878	0.852	440
Affirmed	0.955	0.961	0.951	2930
Macro	0.86	0.875	0.846	3731
Weighted	0.927	0.927	0.929	3731

(a) Presence average 10 fold CV 90/10 ratio

CLS	F1	P	R	Support Test (10% of total)
Family	0.801	0.865	0.751	13
Other	0.823	0.838	0.809	205
Patient	0.977	0.975	0.98	1649
macro	0.867	0.893	0.847	1867
weighted	0.959	0.959	0.959	1867

(b) Experiencer average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Recent	0.969	0.964	0.94	1655
Past	0.771	0.807	0.74	162
Future	0.667	0.706	0.74	50
macro	0.803	0.825	0.783	1867
weighted	0.943	0.943	0.945	1867

(c) Temporality average 10 fold CV 90/10 ratio

Appendix D.2. Meta Annotation Results Breakdown for SLaM

Aggregate results for each defined meta-annotation at SLaM. Performance is aggregated over all extracted concepts listed in Appendix A. We defined the following meta-annotation tasks:

- Status: is the concept affirmed to be affecting the patient or not, values: [Patient / Other / NA]
- Diagnosis: is the concept a diagnosis related to the patient, or not, values: [Yes, No]

Appendix D.3. Meta Annotation Results Breakdown for UCLH

Aggregate results for each defined meta-annotation at UCLH. Performance is aggregated over all extracted concepts listed in Appendix A. We defined the following meta-annotation tasks:

- Negation: is the concept negated or not, values: [Yes / No]
- Experiencer: is the concept experienced by the patient or not, values: [Patient, Other]
- Problem Temporality: is the concept referring to a historical mention, values [Past Medical Issue, Current Problem]

Table D.4: Meta Annotation Results at SLaM

CLS	F	P	R	Support Test (10% of total)
NA	0.873	0.869	0.878	43
Other	0.544	0.663	0.475	7
Affirmed	0.908	0.893	0.924	60
Macro	0.775	0.812	0.757	109
Weighted	0.873	0.874	0.873	109

(a) Status average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Yes	0.931	0.935	0.926	68
No	0.872	0.889	0.880	39
Macro	0.904	0.908	0.905	109
Weighted	0.913	0.912	0.913	109

(b) Diagnosis average 10 fold CV 90/10 ratio

- Certainty: is the concept confirmed to be present, values: [Confirmed, Suspected]
- Irrelevant: is the concept relevant, values: [Yes, No]

Table D.5: Meta Annotation Results at UCLH

CLS	F	P	R	Support Test (10% of total)
Yes	0.896	0.895	0.900	46
No	0.688	0.767	0.631	394
Macro	0.836	0.767	0.631	440
Weighted	0.970	0.969	0.971	440

(a) Negation: average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Other	0.681	0.883	0.65	3
Patient	0.998	0.997	0.999	437
Macro	0.940	0.940	0.825	440
Weighted	0.996	0.996	0.996	440

(b) Experiencer: average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Past Medical Issue	0.710	0.758	0.676	23
Current Problem	0.985	0.981	0.988	412
Macro	0.848	0.870	0.832	435
Weighted	0.970	0.969	0.971	435

(c) Problem Temporality: average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Confirmed	0.985	0.980	0.989	395
Suspected	0.688	0.767	0.631	21
Macro	0.836	0.874	0.810	416
Weighted	0.970	0.970	0.971	416

(d) Certainty: average 10 fold CV 90/10 ratio

CLS	F	P	R	Support Test (10% of total)
Yes	0.685	0.846	0.579	24
No	0.986	0.976	0.994	415
Macro	0.835	0.911	0.787	439
Weighted	0.969	0.970	0.972	439

(e) Irrelevant: average 10 fold CV 90/10 ratio

References

[1] R. Jackson, I. Kartoglu, C. Stringer, G. Gorrell, A. Roberts, X. Song, H. Wu, A. Agrawal, K. Lui, T. Groza, D. Lewsley, D. Northwood, A. Folarin, R. Stewart, R. Dobson, CogStack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital, BMC Med. Inform. Decis. Mak. 18 (1) (2018) 47. doi:10.1186/s12911-018-0623-9.

[2] M. Q. Stearns, C. Price, K. A. Spackman, A. Y. Wang, SNOMED clinical terms: overview of the development process and project status, Proc. AMIA Symp. (2001) 662–666.

[3] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (Database issue) (2004) D267–70. doi:10.1093/nar/gkh061.

[4] T. Searle, Z. Kraljevic, R. Bendayan, D. Bean, R. Dobson, MedCATTrainer: A biomedical free text annotation interface with active learning and research use case specific customisation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 139–144. doi:10.18653/v1/D19-3024.

[5] D. M. Bean, Z. Kraljevic, T. Searle, R. Bendayan, G. Kevin O, A. Pickles, A. Folarin, L. Roguski, K. Noor, A. Shek, R. Zakeri, A. M. Shah, J. T. Teo, R. J. Dobson, ACE-inhibitors and angiotensin-2 receptor blockers are not associated with severe SARS-COVID19 infection in a multi-site UK acute hospital trust, Eur. J. Heart Fail. (Jun. 2020). doi:10.1002/ejhf.1924.

[6] E. Carr, R. Bendayan, D. Bean, M. Stammers, W. Wang, H. Zhang, T. Searle, Z. Kraljevic, A. Shek, H. T. T. Phan, W. Muruet, A. J. Shinton, T. Shi, X. Zhang, A. Pickles, D. Stahl, R. Zakeri, K. O'Gallagher, A. Folarin, L. Roguski, F. Borca, J. Batchelor, X. Wu, J. Sun, A. Pinto, B. Guthrie, C. Breen, A. Douiri, H. Wu, V. Curcin, J. T. Teo, A. Shah, R. Dobson, Evaluation and improvement of the national early warning score (NEWS2) for COVID-19: a multi-hospital study (Apr. 2020). doi:10.1101/2020.04.24.20078006.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[9] J. Howard, S. Ruder, Universal language model fine-tuning for text classification (Jan. 2018). arXiv:1801.06146.

[10] C. Jonquet, N. H. Shah, M. A. Musen, The open biomedical annotator, Summit Transl Bioinform 2009 (2009) 56–60.

[11] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: An unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094. doi:10.1038/srep26094.

[12] I. Landi, B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, R. Miotto, Deep representation learning of electronic health records to unlock patient stratification at scale, NPJ Digit Med 3 (2020) 96. doi:10.1038/s41746-020-0301-z.

[13] A. R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236. doi:10.1136/jamia.2009.002733.

[14] G. Gorrell, X. Song, A. Roberts, Bio-YODIE: A named entity linking system for biomedical text (Nov. 2018). arXiv:1811.04860.

[15] H. Wu, G. Toti, K. I. Morley, Z. M. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale, G. Gorrell, A. Roberts, M. Broadbent, R. Stewart, R. J. B. Dobson, SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research, J. Am. Med. Inform. Assoc. 25 (5) (2018) 530–537. doi:10.1093/jamia/ocx160.

[16] P. J. Gorinski, H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, B. Alex, Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches (Mar. 2019). arXiv:1903.03985.

[17] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5)

(2010) 507–513. doi:10.1136/jamia.2009.001560.

[18] D. Ferrucci, A. Lally, UIMA: an architectural approach to unstructured information processing in the corporate research environment, Natural Language Engineering (2004) 1–26.

[19] T. Morton, J. Kottmann, J. Baldridge, G. Bierner, Opennlp: A java-based nlp toolkit, in: Proc EACL, 2005.

[20] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and robust models for biomedical natural language processing (Feb. 2019). arXiv:1902.07669.

[21] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu, Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines, Journal of the American Medical Informatics Association (2017) ocx132doi:10.1093/jamia/ocx132.

[22] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic Acids Res. 39 (Web Server issue) (2011) W541–5. doi:10.1093/nar/gkr469.

[23] A. Keselman, C. A. Smith, G. Divita, H. Kim, A. C. Browne, G. Leroy, Q. Zeng-Treitler, Consumer health concepts that do not map to the UMLS: Where do they fit?, J. Am. Med. Inform. Assoc. 15 (4) (2008) 496–505. doi:10.1197/jamia.M2599.

[24] K. C. Wang, Standard lexicons, coding systems and ontologies for interoperability and semantic computation in imaging, J. Digit. Imaging 31 (3) (2018) 353–360. doi:10.1007/s10278-018-0069-8.

[25] Data protection and information governance, https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/, accessed: 2020-7-31.

[26] J. Hellrich, U. Hahn, Fostering multilinguality in the UMLS: A computational approach to terminology expansion for multiple languages, AMIA Annu. Symp. Proc. 2014 (2014) 655–660d.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 3111–3119.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding (Oct. 2018). arXiv:1810.04805.

[29] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. doi:10.1162/tacl_a_00051.

[30] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[31] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Sci Data 3 (2016) 160035. doi:10.1038/sdata.2016.35.

[32] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, J. Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition, Bioinformatics 34 (8) (2018) 1381–1388. doi:10.1093/bioinformatics/btx761.

[33] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, J. Han, Cross-type biomedical named entity recognition with deep multi-task learning, Bioinformatics 35 (10) (2019) 1745–1752. doi:10.1093/bioinformatics/bty869.

[34] B. Xu, X. Shi, Z. Zhao, W. Zheng, Leveraging biomedical resources in Bi-LSTM for Drug-Drug interaction extraction, IEEE Access 6 (2018) 33432–33439. doi:10.1109/ACCESS.2018.2845840.

[35] A. Mascio, Z. Kraljevic, D. Bean, R. Dobson, R. Stewart, R. Bendayan, A. Roberts, Comparative analysis of text classification approaches in electronic health records, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020, pp. 86–94. doi:10.18653/v1/2020.bionlp-1.9.

[36] S. Mohan, D. Li, MedMentions: A large biomedical corpus annotated with UMLS concepts (Feb. 2019). arXiv:1902.09476.

[37] D. L. Mowery, S. Velupillai, B. R. South, L. Christensen, D. Martinez, L. Kelly, L. Goeuriot, N. Elhadad, S. Pradhan, G. Savova, Others, Task 2: ShARe/CLEF ehealth evaluation lab 2014, 2014.

[38] O. Bodenreider, A. T. McCray, Exploring semantic groups through visual approaches, J Biomed Inform 36 (6) (2003) 414–432.

[39] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. B. A. McDermott, Publicly available clinical BERT embeddings (Apr. 2019). arXiv:1904.03323.

[40] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. P. Palmer, X. Shi, T. Cai, I. S. Kohane, Clinical concept embeddings learned from massive sources of multimodal medical data (Apr. 2018). arXiv:1804.01486.

[41] R. Zakeri, R. Bendayan, M. Ashworth, D. M. Bean, H. Dodhia, S. Durbaba, K. O Gallagher, C. Palmer, V. Curcin, E. Aitken, W. Bernal, R. D. Barker, S. Norton, M. C. Gulliford, J. T. Teo, J. Galloway, R. J. Dobson, A. M. Shah, A case-control and cohort study to determine the relationship between ethnic background and severe COVID-19 (Jul. 2020). doi:10.1101/2020.07.08.20148965.

[42] K. C. Fraser, I. Nejadgholi, B. De Bruijn, M. Li, A. LaPlante, K. Z. El Abidine, Extracting UMLS concepts from medical text using general and Domain-Specific deep learning models (Oct. 2019). arXiv:1910.01274.

[43] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (7763) (2019) 95–98. doi:10.1038/s41586-019-1335-8.

[44] M. Krauthammer, G. Nenadic, Term identification in the biomedical literature, J. Biomed. Inform. 37 (6) (2004) 512–526. doi:10.1016/j.jbi.2004.08.004.

Chapter 5: An NLP Approach to Uncover the Incidence and Prevalence of Epilepsy and Associated Comorbid Conditions

5.1 Summary

Introduction

People with epilepsy often have frequent encounters with healthcare providers which makes routinely collected healthcare data an ideal source of information regarding the health of people with epilepsy, in an ecologically valid environment. We aimed to explore the feasibility of using a novel patient data analytics tool (CogStack ecosystem) to identify people with epilepsy from free-text letters generated following neurology clinic attendances at King's College Hospital (KCH), and to automatically extract information on demographic patterns and common comorbidities associated with the different types of epilepsies.

Method

To identify our cohort, we utilised CogStack to search and retrieve all Neurology clinic letters which contained the word "epilepsy" between 2013-2020. We then used a novel AI-based natural language processing system called the Medical Concept Annotation Tool (MedCAT) to annotate the free-text of each of the clinic letters. MedCAT is a semi-supervised system that leverages the SNOMED-CT terminology structure to annotate free-text. Epilepsy specialists were asked to produce training and benchmarking materials for MedCAT through annotating a minimum of 400 documents for all mentions of diseases, symptoms and medications, as well as to label the contextual information of each annotation (Meta-annotation). The trained model of MedCAT was then used to automatically annotate all clinic letters and its performance was evaluated.

Results

CogStack retrieved 36,855 documents relating to 9,860 unique patients, from which MedCAT could accurately identify that 4,011 patients had epilepsy. Epilepsy affected both genders equally across all epilepsy types. Lastly there is an increasing demand on neurology services as people with epilepsy visiting a neurologist is increasing by 4.60% a year.

Conclusions

Overall automated machine-learning based, information extraction techniques tailored to patient records, particularly unstructured data, can provide insights beyond direct patient care but for evidence-based healthcare provider service improvements and decision making.

5.2 Authors

Anthony Shek, Pedro Viana, Elisa Bruno, Chirag Mehra, Samia Elkommos, Joel Winston, James T H Teo, Mark P. Richardson

Statement of Contributions

Conceptualization: AS, JT, MR;
Data curation: AS, PV, EB, CM, SE, JW;
Formal analysis AS;
Supervision: JT, MR;
Writing-review: AS, JT, MR;

5.3 Introduction

Epilepsy is one of the most common serious chronic neurological conditions that affect both men and women of all ages and has a worldwide prevalence (Ngugi et al., 2010). The Global Burden of Disease Study 2016 revealed that there are 45.9 million people who suffer from active epilepsy globally (GBD, 2019).

Epilepsy is not a single disease but rather refers to a class of neurological disorders which cause deleterious functional changes within the brain, specifically excessive and uncontrolled neuronal activity that manifests itself as epileptic seizures. Epilepsy is classified into four main types: 1) focal, 2) generalised, 3) combined generalised and focal, 4) unknown (Scheffer et al., 2017). The 'unknown' epilepsy type is a category in which the seizures are of unknown onset type or the clinician has not yet gathered sufficient clinical information to be certain about the epilepsy classification.

In epilepsy management, seizure control is the primary objective. However, the quality of life (QOL) is affected by comorbid conditions that include neurological, neuropsychiatric, and neurobehavioral disorders. Roughly 50% of adults with active epilepsy have at least one comorbid medical disorder (Keezer et al., 2016). Appreciation of the relevance of these comorbidities is increasing because they affect epilepsy prognosis, hospital service utilisation, and QOL. For example, migraine and psychiatric comorbidities are associated with poorer seizure outcomes, whereas depression has been linked with reduced QOL (Taylor et al., 2011).

People with epilepsy often have frequent encounters with healthcare providers and the condition itself is considered to be ambulatory care sensitive, meaning that high quality care may reduce unnecessary future inpatient and emergency department follow-ups (Grinspan et al., 2018; Wilner et al., 2014). Since each interaction is recorded, the information contained in the electronic patient record (EPR) can help to identify service utilisation patterns or discrepancies in people with epilepsy early on, potentially reducing the long-term cost to the healthcare system. This is especially important when considering that the main driver of the cost of care for people with epilepsy is hospital service utilisation (Taylor et al., 2011). This cost substantially increases when there is co-occurrence of other medical conditions in people with epilepsy. As it has been shown that there is increased risk of readmission and generally incur medical costs almost 1.4 times higher when compared to those without such comorbidities (Lee et al., 2005).

This makes natural language processing, specifically the task of automated extraction of clinical concepts from EPR, highly attractive to aggregate data of patients from an ecologically valid environment. Numerous rule-based NLP approaches have already been explored such as the extraction of epilepsy clinical text (ExECT) system using the general architecture for text engineering (GATE) framework (Fonferko-Shadrach et al., 2019), the use of regular expressions (REGEX) for the extraction of risk factors for sudden unexpected death in epilepsy (Barbour et al., 2019), or the clinical Text Analysis and Knowledge Extraction System (cTAKES) to exclude psychogenic nonepileptic seizure false positives in an epilepsy cohort (Hamid et al., 2013).

All of these studies have their limitations due to the fact that they rely upon keywords and defined language patterns. They require predefined extraction rules to query the unstructured text for each clinical concept of interest. However, for an accurate retrieval of results, these rules must encompass all possible alternative nomenclature, spelling mistakes, and acronyms, whilst taking into consideration the context in which the concept is used. For example, in an exercise to retrieve all people with epilepsy, a simple query of “epilepsy” across documents that contained mentions such as “Family hx: epilepsy”, or “This pt does not have epilepsy” would result in false positives. Consequently, these rules can become convoluted very quickly, especially during complex queries to retrieve multiple clinical concepts.

5.4 Aims and Objectives

In this project, we use CogStack–MedCAT, a new hospital enterprise search tool which incorporates a machine-learning clinical informatics pipeline to overcome the limitations of rule-based tools (Jackson et al., 2018; Kraljevic et al., 2021). The primary aim is to identify patients with epilepsy from neurology clinic letters from King’s College Hospital (KCH) and to identify demographic patterns and common comorbidities associated with the different types of epilepsies.

5.5 Methods

5.5.1 Ethics Statement

This project was conducted under London South East Research Ethics Committee (reference 18/LO/2048) approval granted to the King’s Electronic Records Research Interface (KERRI); specific work on epilepsy research was reviewed with expert patient input on a virtual committee with oversight from a senior person responsible for protecting the confidentiality of health and care information (known in the UK NHS as a ‘Caldicott Guardian’). The study adhered to the principles of the UK Data Protection Act 2018, UK National Health Service (NHS) information governance requirements, and the Declaration of Helsinki.

5.5.2 Study Population, Data Extraction and Processing

To identify our patient cohort, we utilised CogStack (Jackson et al., 2018) to search and retrieve all neurology clinic letters which contained the word “epilepsy” between 2013-2020. Patients’ demographics were retrieved and the contents of each document containing the clinical characteristics were automatically extracted using a novel AI-based natural language processing system called the MEDical Concept Annotation Tool (MedCAT) (Kraljevic et al., 2021). MedCAT is a semi-supervised machine-learning system that extracts concepts from unstructured text and maps them onto a standardised clinical vocabulary, SNOMED CT.

After all relevant documents were retrieved, a stratified sample of 400 documents by patient age, gender and year of document were used to fine-tune the model, which had previously been provided unsupervised training for SNOMED-CT concepts across ~18million documents in a process outlined in Kraljevic et al. 2021. This supervised training step involved epilepsy specialists (PV, EB, CM, SE) who were asked to annotate a subsample of the free-text documents to produce training and benchmarking materials for MedCAT. An annotation guideline had been written for the annotators to follow and to help standardise the interpretation of the annotation task (see supplementary material item 1). Any disagreements or discrepancies in annotations were discussed and resolved by a third reviewer, AS (technical questions) or MR (clinical questions).

The annotators each annotated a minimum of 400 neurology clinic letters for all mentions of diseases, symptoms, and medications. An example of how annotations are presented to the reviewer is displayed in figure 11.

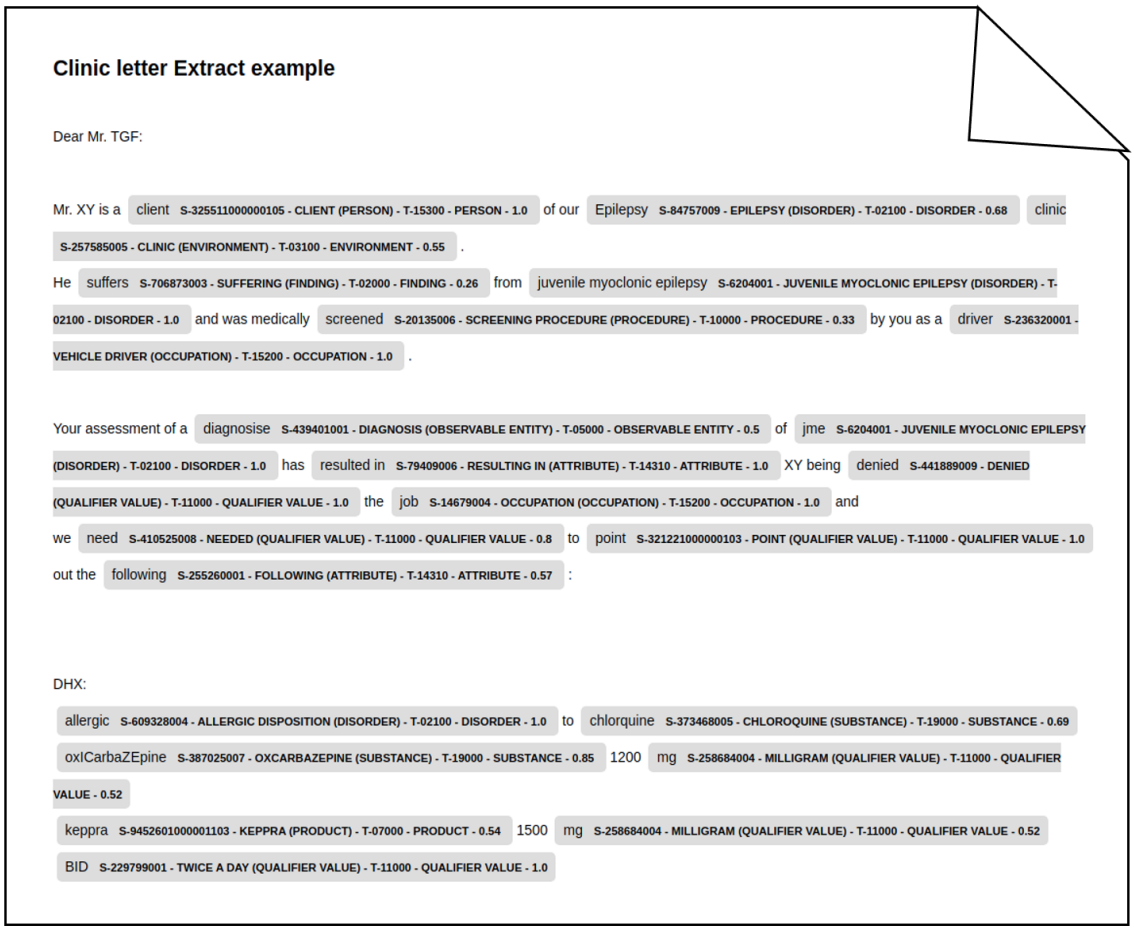


Figure 11. Demonstration of how the MedCAT toolkit displays annotations in a neurology clinic letter. Each grey box represents a model predicted annotation with linked meta-data in bold. No SCTID concept filters were applied in the example to demonstrate the breadth of knowledge possible to be explored with this tool.

Concurrent with the concept labelling, annotators were instructed to label the contextual information of each annotation (meta-annotation). These meta annotations included: Subject/Experiencer with options [“Patient”, “Family”, “Other”], Presence of concept [“True”, “False”, “Hypothetical or n/a”], and Time relative to admission [“Past”, “Present”, “Future”]. Once trained, the MedCAT model was used to annotate all relevant SNOMED concepts in the selected neurology clinic letters at KCH. Only the concepts which had the following contextual labels (meta annotations) were extracted:

- Subject/Experiencer: “Patient” and
- Presence: “True” and
- Time: “Past” or “Recent”

Snippets of this process are demonstrated in figure 12.

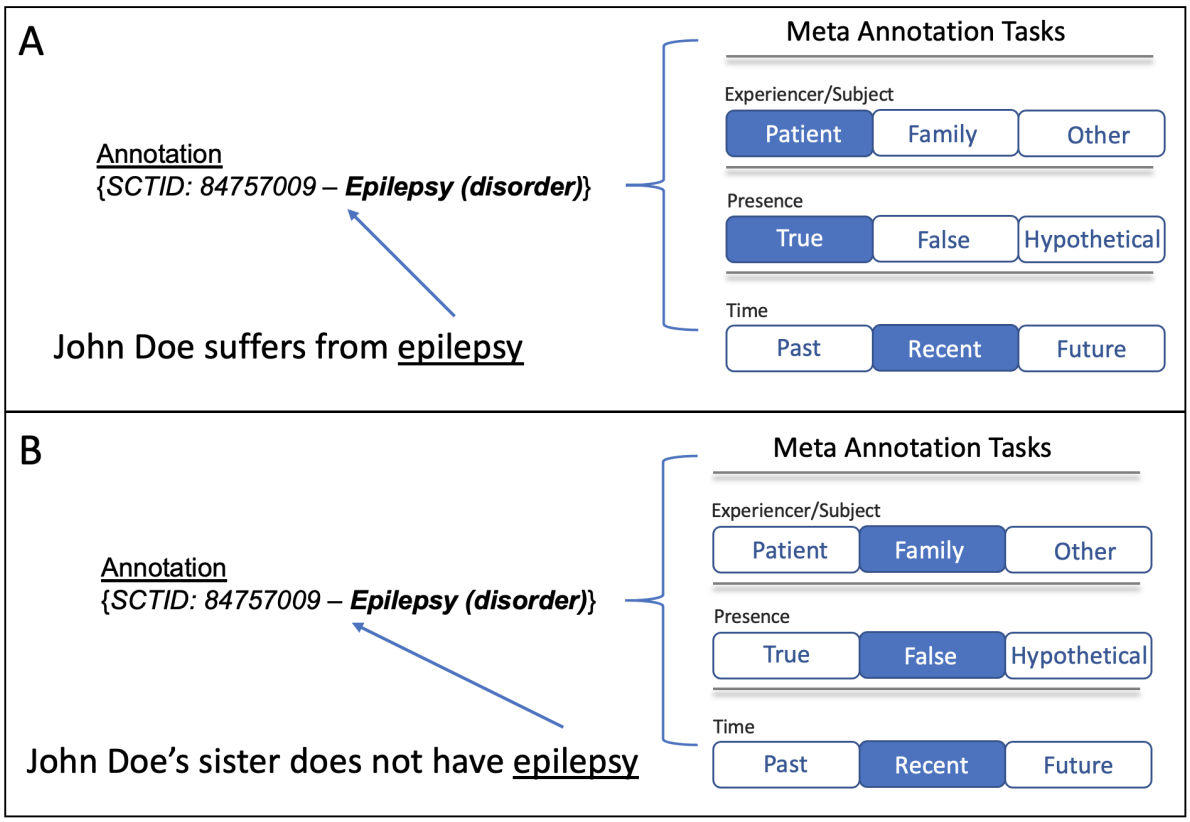


Figure 12. Representation of the meta annotation tasks and how different option combinations can capture the contextual representation of the annotation. Two sentences containing the concept 'Epilepsy' are provided. Panel A represents a combination of meta-annotations that state that the patient currently has epilepsy. Panel B displays a different selection of options which reflect an 'Epilepsy' mention in a different context.

Once a labelled training set was created, a MedCAT model was trained. The MedCAT model extracts clinical information as an annotation and links it to any associated meta-data in a JSON format file (JavaScript Object Notation). JSON is an open standard file format and data interchange format that consists of attribute–value pairs and arrays. An example of the annotation and the meta-data stored can be seen in figure 13. The annotation would include: the recognised SNOMED-CT concept, source value, name of the annotator, start and end position of the label created, time of annotation and meta-annotation values.

“The patient does not suffered with ***epielepsy***”


```
"annotations": [{"id": 1,
                 "user": "admin",
                 "sctid": "84757009",
                 "name": ”Epilepsy (disorder)",
                 "source_value": "epielepsy",
                 "start": 35, "end": 44,
                 "validated": True,
                 "correct": True,
                 "deleted": false,
                 "alternative": false,
                 "terminated": false,
                 "last_modified": "2021-08-25 17:23:41.557191+00:00",
                 "manually_created": false,
                 "acc": 0.69,
                 "meta_anns": [{"Experiencer": Patient},
                               {"Presence": False},
                               {"Time": Recent}]}]
```


Figure 13. Example of the inference of structure from unstructured text through an annotation. In this example a misspelt disorder “epielepsy” is linked to the SNOMED concept “Epilepsy (disorder)” and it is currently, not present in the patient.

5.5.3 SNOMED–CT

The edition of SNOMED-CT used in this project was the 2020-10-01 UK extension combined with the 2020-10-01 UK drugs extension. A list of relevant SNOMED-CT target concepts was selected which included terms for epilepsy and several common comorbidities. For each target concept, all children concepts as defined by the inbuilt SNOMED-CT “IS A” hierarchical relationship structure were also extracted and counted towards the corresponding target concept. For example, if the concept “juvenile myoclonic epilepsy (disorder)” was detected this would be counted towards the target term “epilepsy (disorder)”. The full list of extracted target concepts and respective number of child concepts can be found in supplementary material item 2A.

5.5.4 Reorganisation of Mapping of Epilepsy Terms to ILAE Epilepsy Classification

The version of SNOMED-CT structure does not reflect the current ILAE classifications system. Therefore, concepts were re-mapped to an appropriate heading to be grouped to the appropriate epilepsy classification structure. This was only done for all children concepts under the SNOMED term “Epilepsy (disorder)” SCTID: 84757009. They were reclassified into epilepsy type. This work was conducted by the epilepsy specialists on

the investigator team, PV, JW, MR. Please see supplementary material 4 for the full reorganisation of epilepsy-related SNOMED concepts.

An outline of the entire annotation and extraction process is shown in figure 14.

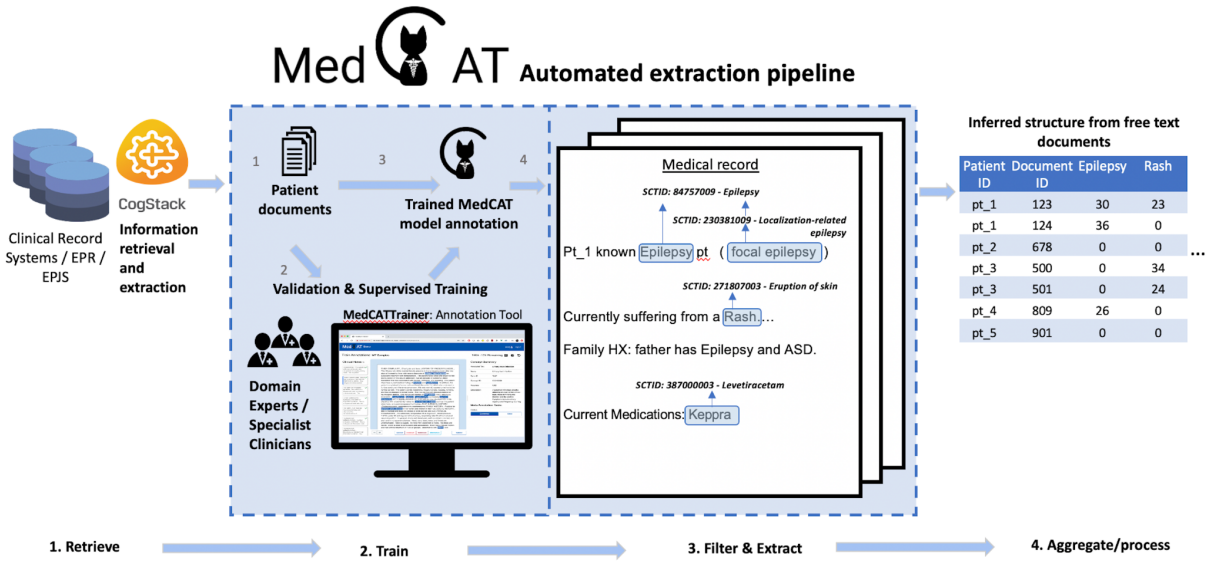


Figure 14. The outline of the MedCAT pipeline. Step 1: Retrieve the data. Step 2: Train the data extraction model on a subset of the dataset with domain specific experts to validate the model. Step 3: Once Trained run the model on the entire dataset to filter and extract terms in the context of the patient. Step 4: Aggregate and process the data.

5.5.5 Data and Statistical Analyses

All data and statistical analyses were performed in python 3.8 using the python packages: medcat v0.4, sklearn v0.24, scipy v1.7, and pandas v1.3. Data visualisations were created using matplotlib v3.4, and plotly v5.1.

5.6 Results

5.6.1 Document Retrieval

Cogstack retrieved 36,855 documents from 9,860 unique patients who attended a neurology clinic between 2013-2020. Despite all documents containing the word “epilepsy”, MedCAT identified that only 4,011 patients had a diagnosis of epilepsy across 25,291 documents. The patients who were removed had no mention of an epilepsy diagnosis within the 7-year period and their documents only contained phrases such as: “Seen in Epilepsy Clinic”, “Seen by Epilepsy Nurse”, “signed by Dr XXX Epilepsy consultant”.

5.6.2 Data Extraction

The performance of MedCAT is measured and broken down into two categories. 1) The performance of identifying the annotations itself and 2) the identification of the contextual information as represented through the Meta-annotations.

Annotation Performance

Once a model was trained, the model's annotations were benchmarked against an unexposed sample of annotations (not seen by the model) created by the specialised epilepsy clinicians. A sample of grouped concept extraction performances is displayed in table 1. The full list of grouped concept performances can be found in the supplementary material item 3A.

Concept	Count	Precision	Recall	F1-Score
Epilepsy- Total	604	0.971	0.876	0.921
Epilepsy type - Unknown	522	0.983	0.866	0.921
Epilepsy type - Focal	36	0.895	0.944	0.919
Epilepsy type - Generalised	26	0.885	0.885	0.885
Epilepsy type - Combined generalised and focal	20	1.000	1.000	1.000
Diabetes	372	0.894	0.745	0.812
Anxiety	31	0.867	0.839	0.852
Cerebrovascular accident	418	0.939	0.847	0.891
Dementia	68	0.943	0.971	0.957
Asthma	63	0.332	0.984	0.496
Myocardial infarction	43	0.932	0.953	0.943
Chronic obstructive lung disorder	384	0.888	0.891	0.889

Table 1. Selected list of Annotation performance by SNOMED concept group. The full list of concept performances is displayed in the supplementary materials 3A.

Meta-annotation Performance

Table 2 displayed the average meta-annotation performance across all concept annotations. The F1 is evaluated on a scale ranging from 0 to 1, where 1 represents that the model has perfectly classified each annotation into the correct class.

Meta annotation	Precision	Recall	F1-Score
Subject/Experiencer	0.987	0.987	0.977
Presence	0.940	0.943	0.940
Time	0.959	0.961	0.956

Table 2. The meta-annotation performance across the 3 meta-annotation tasks.

Annotator challenges

Several main challenges of the annotation process included: 1) labelling the text with the SNOMED concept which encapsulates what the author meant. For examples “The patient has tonic-clonic epilepsy” was at times interpreted as “generalised epilepsy” and in other times “tonic-clonic seizures”. These nuances at times required highly specialised domain knowledge. 2) Discerning between the meta-annotation tasks: Recent and Past was, at times, challenging. This was because the categorisation of these meta-terms was not always obvious, and annotators needed to discuss its interpretation. When did “recent” become a “past” mention? At times there was not enough information within the document to make a clear decision whether the concept was immediately relevant to this admission episode. 3) The large quantity of SNOMED concepts was difficult for the annotators to remember and label the document with. This would result in the annotator’s mislabelling or omitting the most appropriate concept annotation code.

5.6.3 Epilepsy Classification

The size of our cohort was n=4,011 people with epilepsy, of which there was an equal gender distribution where 50.04% were female. The discrete numbers of patients categorised by epilepsy type and gender is represented in figure 15. The unknown epilepsy type represented the largest category (n=1,738) where female patients represented 46.78%. A diagnosis of focal epilepsy was the largest determined

epilepsy type (n=1,495) where females represented 51.64%. The number of patients that reached a diagnosis of generalised epilepsy was about a third of the size of focal epilepsy (n=622) but had a larger proportion of female patients equating to 56.27%. The smallest category was the combined generalised and focal epilepsy type (n=156) and females only made up 46.15% of these patients.

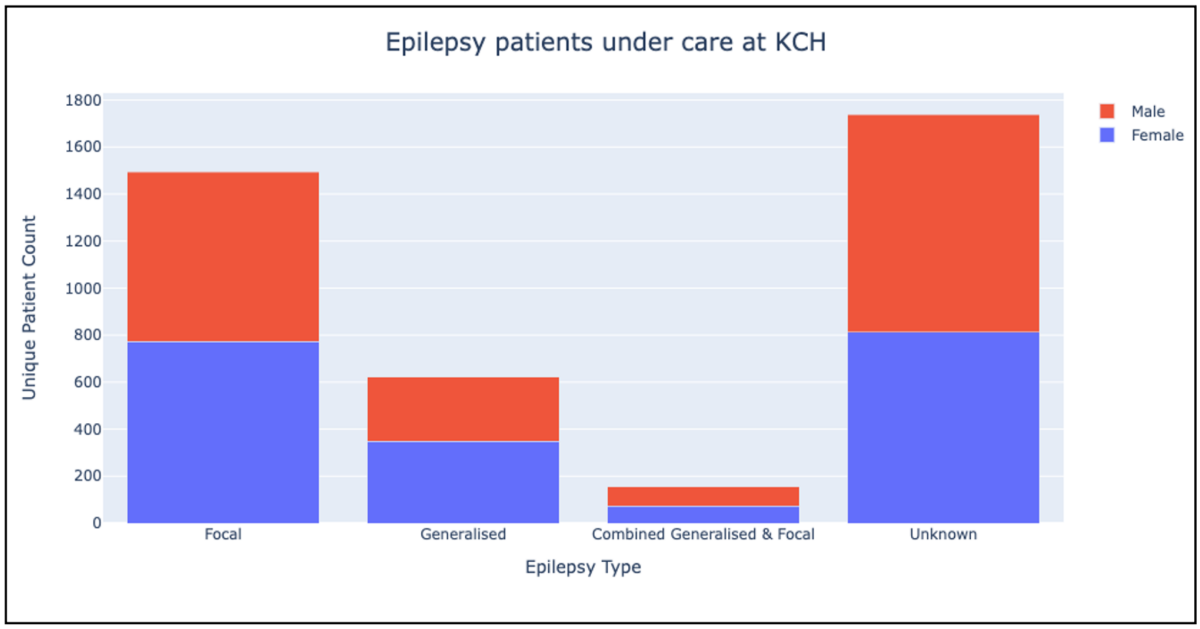


Figure 15. The overall number of people with epilepsy seen across a 7-year period. Categorised by epilepsy type and gender.

5.6.4 Outpatient Neurology Service Utilisation

The number of people with epilepsy attending outpatient neurology clinics was increasing each year between 2013 and 2019 as shown in figure 16. Across the 7-year time period of this study, the number of people with epilepsy using outpatient neurology service demand increased by 30.79%. The average annual increase was 4.60%.

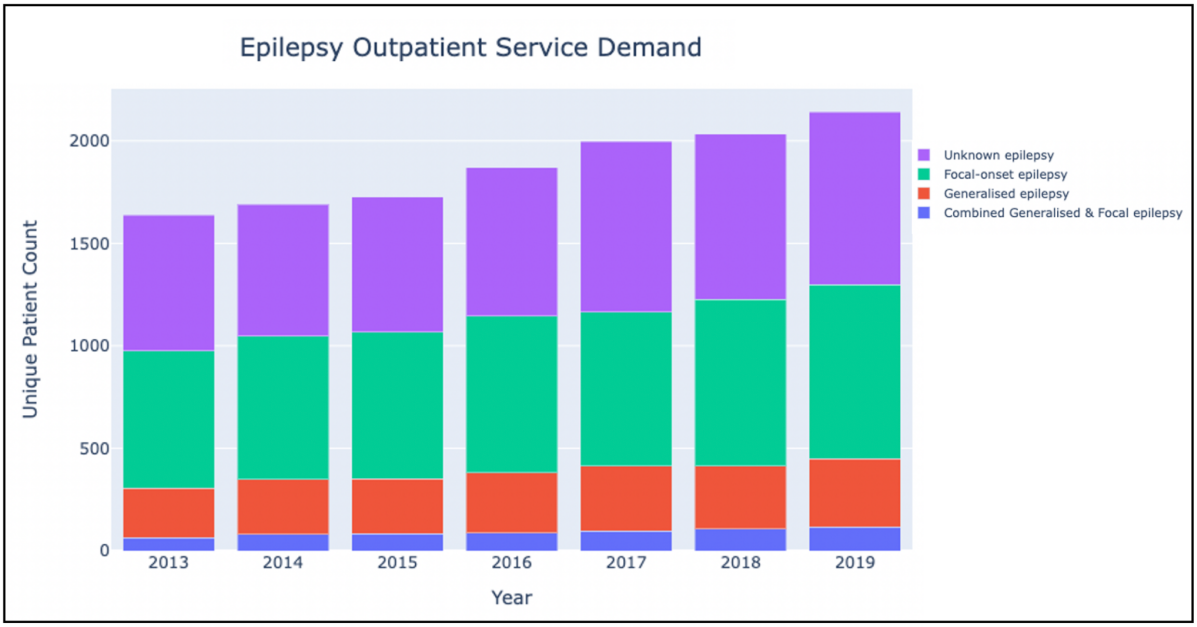


Figure 16. The increasing demand for outpatient epilepsy services across time.

5.6.5 Comorbid Conditions

A number of comorbid conditions were selected to broadly represent psychiatric, endocrine/metabolic, cardiovascular, and respiratory disorders. A breakdown of the prevalence of each comorbid condition investigated and epilepsy type is displayed in figure 17. Psychiatric disorders were by far the most common comorbidity across all epilepsy types.

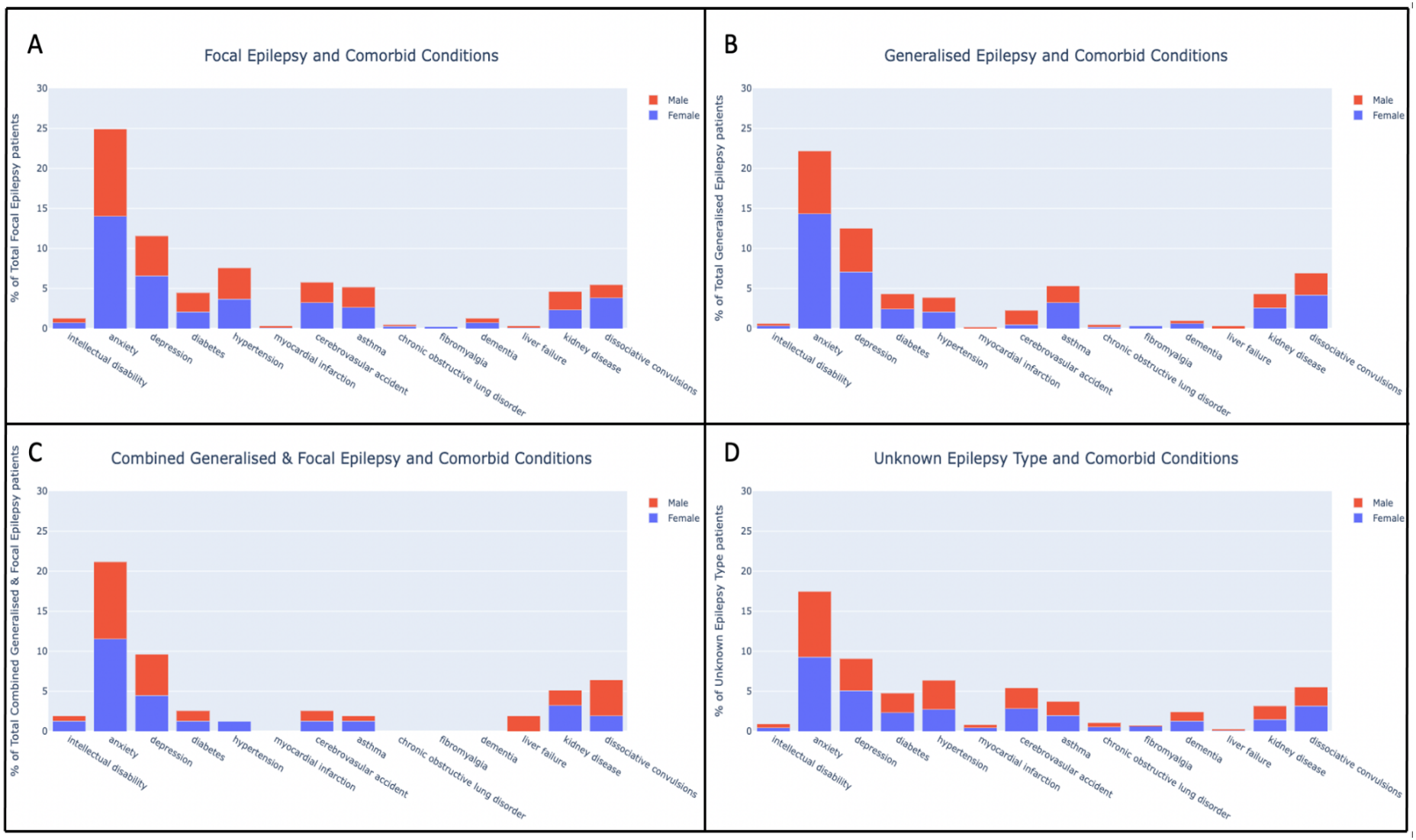


Figure 17. The prevalence of comorbid conditions across epilepsy types.

5.7 Discussion

5.7.1 Annotation Tools and Previous Work

NLP is increasingly being looked at as a tool for automating clinical information extraction purposes. Previous epilepsy-related retrospective studies have used rule-based NLP tools such as JAPE and REGEX (Barbour et al., 2019; Fonferko-Shadrach et al., 2019; Hamid et al., 2013). These are ideal for extracting semi-structured data such as clinic dates, gender, and dates of birth. However, these techniques are limited by the fact that it can be difficult to account for the variability of the language used to express clinical information. For larger extractions, identifying language patterns can become very complex with potentially conflicting rules. Therefore, writing generalisable queries that extract every item of information is not feasible as a solution to ensure exhaustive text extraction. We have demonstrated the potential of a pipeline which includes a new machine-learning NLP approach to accurately extract a wide variety of epilepsy-related information. This machine learning tool has the advantage of not requiring prespecification of language patterns, and allows for context-dependent extraction. It leverages the structure of standardised terminologies, in this case SNOMED-CT to appropriately standardise medical language into unique concepts and group related concepts into a hierarchical knowledge graph. Future clinical NLP applications will have a hybrid NLP-pipeline which consists of both rule-based methods to complement a machine learning approach.

5.7.2 Epilepsy Types

In our study we found that 4,011 patients with a diagnosis of epilepsy had an outpatient appointment at KCH between 2013-2020. During this period only slightly more than half of people with epilepsy (56%) were assigned to one of the diagnostic categories of the ILAE (Scheffer et al., 2017). This may suggest that 1) many patients are still undergoing investigations in determining epilepsy type, 2) there is an underuse of explicit mentions of an epilepsy type during health documentation, 3) despite thorough investigation, insufficient information could be identified to allow classification (e.g. there is no one available to provide a witness history, or an MRI cannot be performed). In cases 1 & 2, it may be possible to extract seizure type which can then be mapped to the epilepsy type. Broadly, patients with focal seizures have focal epilepsies,

patients with generalised seizures have generalised epilepsies, patients with focal and generalised seizures have focal and generalised epilepsies. However in practice until an epilepsy type is established, patients who appear to have generalised seizures may actually have a focal epilepsy i.e. the seizures are actually 'secondary generalised' (AKA ‘focal to bilateral tonic-clonic seizures') and in these cases the 1-to-1 mapping of seizures type to epilepsy type would lead to the incorrect classification. Whilst this study has used all epilepsy subtype diagnosis concepts available within SNOMED-CT and mapped them to the appropriate epilepsy type, future investigations should extract and map SNOMED-CT seizure concepts into the corresponding ILAE seizure onset classification structure. This will help to gather further insights into those with a not yet known epilepsy-type, but a predominant seizure type. Large scale information extraction via NLP might contribute by identifying the gaps in cases (1&2) - i.e. identifying that elements of history or investigations have not been done yet, and prompting the clinician to fill the gaps. Additionally, it might contribute to case (3) (and cases (1&2)) by showing that the available information clusters very close to a definite epilepsy type or subtype (or perhaps is compatible with only a single category), despite gaps in obtaining the required information.

In our study, of those who received an epilepsy type diagnosis, the distribution showed that focal epilepsy was more common than generalised epilepsy, whilst combined generalised and focal epilepsy was the least prevalent. Whilst some studies support this distribution (Guekht et al., 2010), others found that generalised epilepsy is more common (Carlson et al., 2014; Fiest et al., 2017). The difference in distribution compared to larger-scale epidemiological studies is likely to be a population-specific feature relevant to the nature of the clinical service in a tertiary hospital that tended to focus on focal epilepsies as surgical targets or needing complex characterisation.

In terms of gender differences, at the broad overview of an epilepsy diagnosis, there were no overall differences in our clinical population. This is in accordance with findings from a recent meta-analysis of international studies (Fiest et al., 2017). We further explored gender differences in the epilepsy sub-categories but we also observed no major gender differences in this regard. Our study only looked at the broad level categories across the epilepsy spectrum, however, this limits the view of gender differences at more specific subcategories, for example: Generalised

Epilepsies encompasses Idiopathic Generalised Epilepsies which encompasses the 4 well established epilepsy syndromes: Childhood Absence Epilepsy, Juvenile Absence Epilepsy, Juvenile Myoclonic Epilepsy, and Generalised Tonic–Clonic Seizures Alone. Further subcategorisation at more granular levels of the epilepsy classification, such as epilepsy syndromes, could be explored as it is possible that opposing gender differences within subcategories could cancel out any observed difference.

5.7.3 Comorbidities

Epilepsy is generally regarded as a spectrum of brain conditions characterised by recurrent seizures caused by abnormal paroxysmal neuronal activity. Identification of similar abnormal neuronal patterns or their emergent clinical phenotype allows for the grouping and formation of defined subtype categories. Each of these epilepsy type categories will share a similar predisposition towards comparable paroxysmal activity patterns which could be deemed a common risk factor to some comorbid conditions such as intellectual disability in West Syndrome (Pavone et al., 2020). The comorbidities of epilepsy form the core of these associated conditions and contribute to our evolving conceptualisation of epilepsy as a condition consisting of more than seizures (Berg, 2011; Keezer et al., 2016). On the contrary, many comorbidities may be due to factors that are generalised to many (sometimes all) epilepsy types and not dependent on neuronal firing patterns. For example, anxiety and depression may be due to having an unpredictable long-term condition, and might have exactly the same causative mechanism as in e.g., someone with bad asthma.

In our study, anxiety and depression were the most prevalent across all types of epilepsy. However, due to the limitations of our tools this does not provide any information about the severity of a clinical diagnosis. Another prominent observation was that there were no males, who were suffering from combined generalised and focal epilepsy, that presented with hypertension. One hypothesis could be that male patients with hypertension remain under the "Unknown" epilepsy type, and pass away from sudden unexpected death in epilepsy (SUDEP) before they receive a combined epilepsy type diagnosis. Historically, hypertension has not been considered a major risk factor for SUDEP, however, a pathomechanism has been proposed involving the interplay between the renin-angiotensin system and sympathetic nervous system,

both known to be involved in the development of hypertension and cardiovascular disease, as potentially one of the underlying mechanisms of SUDEP (Szczurkowska et al., 2021). Further studies should investigate this association.

This study is limited as it does not infer any degree of causation for these comorbidities but rather correlations or associations. Straightforward measurement of associations between different diseases has been referred to as an "empirical statistical phenomenon that has no meaning in itself" (Rutter, 1997). However, measurement of associations is an initial step in a process which in turn is followed by investigations into why certain conditions have an increased risk to co-occur.

The UK's ageing population has been associated with an increased proportion of the population with comorbidity (Kingston et al., 2018). Although this study did not investigate comorbidities in the context of age. The incidence of many of the comorbidities explored in our study are known to increase in the elderly, especially ageing-related epileptogenic conditions such as stroke and dementia (Keezer et al., 2016; Kingston et al., 2018).

It is known that psychiatric, endocrine/metabolic, cardiac, and respiratory disorders are associated with worse long-term health outcomes (Giussani et al., 2021; Seidenberg et al., 2009). An examination of USA private insurance claims data showed that in people with epilepsy 80% of direct medical costs were not related to epilepsy, but were related to the treatment of comorbid somatic and psychiatric conditions (Ivanova et al., 2010). Like epilepsy aetiology, it is important that the presence of comorbidities be considered for every patient with epilepsy at each stage of classification, enabling early identification, diagnosis, and appropriate management (Keezer et al., 2016).

5.7.4 Monitoring Service Demand

Data on the financial burden of epilepsy are needed for health-care planning and resource allocation. Our study has shown that there is an increase in the number of people with epilepsy attending neurology clinics (4.60% average annual growth rate over the studied 7-year period). This number only includes whether a patient has

visited a clinic or not in any given year (0 or 1 per year); however, patients are likely to be provided and attend more than one appointment per year and is therefore an underestimate of the overall increase in healthcare service burden.

Overall, this study has shown that taking advantage of routinely collected data from EPR helps to better understand how various factors, including epilepsy subtypes, gender and comorbidities, interplay and that future studies should further investigate these factors with the aim to improve healthcare pathways as well as establish a value based healthcare system.

5.8 Conclusion

The development of automated extraction tools will significantly empower healthcare providers and researchers to take full advantage of all data within EPRs. Our study has demonstrated an accurate automated extraction for epilepsy related concepts using a standardised clinical terminology and has the potential to replace the manual creation of data warehouses.

Our study identified 4,011 patients with epilepsy of which half have not received an epilepsy diagnosis. Epilepsy affected both genders equally across all epilepsy types. Lastly, there is an increasing demand on neurology services as people with epilepsy visiting a neurologist increased on average by 4.60% a year.

Overall, automated information extraction tools applied to routinely collected data can not only be used for service demand monitoring and improvement but to also enhance epilepsy research.

Chapter 6: Anti-seizure Medication Choice and Switching

6.1. Summary

Introduction

Achieving seizure freedom is the primary goal of the treatment and management of epilepsy. Anti-seizure medications (ASMs) are the first line of treatment for new-onset epilepsy. However, although there is evidence to inform the choice of first treatment in epilepsy (A. Marson et al., 2021a, 2021b; A. G. Marson, Al-Kharusi, et al., 2007a, 2007b; A. G. Marson, Appleton, et al., 2007), there is limited evidence to inform the optimal order of subsequent ASMs after the first, or to inform medication combinations since the permutations are too numerous to trial. Real-world data regarding ASM prescribing patterns and patient outcomes might provide relevant evidence. We aimed to explore the feasibility of using a novel patient data analytics tool (CogStack ecosystem) to automatically extract ASM prescribing patterns alongside diseases and symptoms at King's College Hospital NHS Foundation Trust (KCH).

Method

To identify our patient cohort, we utilised CogStack to search and retrieve all Neurology clinic letters which contained the word “epilepsy” between 2013-2020. We then used a novel AI-based natural language processing system called the Medical Concept Annotation Tool (MedCAT) to annotate the free-text of each of the patients with epilepsy neurology clinic letters. MedCAT is a semi-supervised system that leverages the SNOMED-CT terminology structure to annotate free-text. Epilepsy specialists were asked to produce training and benchmarking materials for MedCAT through annotating a minimum of 400 documents for all mentions of diseases, symptoms and medications as well as to label the contextual information of each annotation (meta-annotation). The trained model of MedCAT was then used to annotate all clinic letters and its performance was evaluated.

Results

CogStack retrieved 36,855 documents of 9,860 unique patients. From those documents, MedCAT could accurately identify that 4,011 patients had epilepsy, 26.6%

were seizure-free. Levetiracetam followed by Lamotrigine were the most popular ASMs. Polytherapy prescribing patterns were compared in the context of reaching seizure freedom.

Conclusions

Overall, automated information extraction techniques of patient records, particularly unstructured data, can provide insights beyond direct patient care and into research on treatment strategies and adverse events. Together, this can improve the management of epilepsies through evidence-based decisions.

6.2 Authors

Anthony Shek, Pedro Viana, Elisa Bruno, Chirag Mehra, Samia Elkommos, Joel Winston, James T H Teo, Mark P. Richardson

Statement of Contributions

Conceptualization: AS, JT, MR;
Data curation: AS, PV, EB, CM, SE, JW;
Formal analysis AS;
Supervision: JT, MR;
Writing-review: AS, JT, MR;

6.3 Introduction

Anti-seizure medications (ASMs), previously known as anti-epileptic drugs (AEDs), are the first line of treatment for new-onset epilepsy. Their ultimate goal is to help people with epilepsy achieve seizure-freedom. However, despite there currently being over 20 ASMs, 30% of people with epilepsy do not respond to first line ASMs, and the lack of response is difficult to predict (WHO, 2006). All ASMs are associated with side effects which can vary from one ASM to another (P. Perucca & Gilliam, 2012). Therefore, in practice, the goal of the treatment and management of epilepsy is to maximise seizure control whilst minimising adverse drug effects.

Patients with refractory epilepsy may take 2, 3 or even 4 ASMs in combination in an attempt to control their seizures. However, polypharmacy has been associated with worse quality of life (QOL) in patients with epilepsy despite seizure control (Alexander et al., 2018). Polypharmacy increases the potential for negative medication effects and drug interactions, amongst which neurobehavioural adverse effects are common (Alexander et al., 2018). The practice of initiating new medications to manage the adverse effects of other drugs – the so-called "prescribing cascade effect" (Rochon & Gurwitz, 1997) – is common, with physicians attributing behavioural disruption to other factors, including the epilepsy itself (Plevin et al., 2018).

It can be challenging to find the optimal regime early on and the duration of ASM therapy is often over long periods of time. Additionally, it is possible that patients with years of uncontrolled seizures may not have tried all potentially beneficial ASMs or ASM combinations. There is little evidence for the optimal ASM order or medication combinations since the possibilities are so numerous that it is impossible for doctors and their patients to try every permutation in a single lifetime. The best estimates of the pattern of ASM utilisation for epilepsy have been calculated from the overall prescription of ASMs in primary care. However, this would have included prescriptions for other conditions such as pain, migraine, and psychiatric conditions (Moran et al., 2004; O'Dwyer et al., 2018).

To complicate matters further, roughly 50% of adults with active epilepsy have at least one comorbid medical disorder (Keezer et al., 2016). Appreciation of the relevance of these comorbidities is increasing because they affect epilepsy prognosis and QOL. For example, migraine and psychiatric comorbidities are associated with poor seizure outcome, whereas depression has been linked with reduced QOL (Taylor et al., 2011).

Together, polypharmacy, extended periods of therapy, and comorbidities compound the complexities faced when conducting large-scale research investigating drug efficacy and outcomes. Electronic patient records (EPRs), documenting patients' medical and drug history over time, have the potential to overcome some of these challenges and aid our understanding of anti-epileptic drugs, with the ultimate goal to improve quality of life and even achieve seizure freedom. These sources of information have not been used to their maximum research potential because they are

unstructured and difficult to manage; however, recent advancements in Natural Language processing (NLP) allow for hospitals to take control of their own data.

6.4 Aims

Our aim was to (i) explore the feasibility of using the analytics techniques described in previous chapters to automatically extract ASM prescribing patterns alongside associated symptoms at King's College Hospital NHS Foundation Trust (KCH), (ii) examine prescribing patterns of ASMs in patients with epilepsy, (iii) identify rate of seizure freedom with different ASMs/combinations, and (iv) identify rates of common idiosyncratic side effects/associated symptoms with the top 10 ASM prescribing patterns.

6.5 Methods

6.5.1 Ethics Statement

This project was conducted under London South East Research Ethics Committee (reference 18/LO/2048) approval granted to the King's Electronic Records Research Interface (KERRI); specific work on Epilepsy research was reviewed with expert patient input on a virtual committee with Caldicott Guardian oversight. The study adhered to the principles of the UK Data Protection Act 2018, UK National Health Service (NHS) information governance requirements, and the Declaration of Helsinki.

6.5.2 Study Population, Data Extraction and Processing

This study followed an identical methodology to the previous results chapter 5.The full details can be found in Section 5.5.

6.5.3 Selection of SNOMED-CT Terms and Modifications

A list of relevant SNOMED-CT target concepts was selected which included terms for epilepsy, seizure freedom, ASMs, and selected side effects concepts. The ASMs included were: Acetazolamide, Brivaracetam, Cannabidiol, gabapentin, lamotrigine, levetiracetam, oxcarbazepine, pregabalin, perampanel, primidone, topiramate,

tiagabine, lacosamide, rufinamide, eslicarbazepine acetate, ethosuximide, zonisamide, carbamazepine, clobazam, clonazepam, ethosuximide, phenobarbital, phenytoin, sodium valproate, magnesium valproate, valpromide, primidone, vigabatrin, retigabine.

For each target concept, all children concepts as defined by the inbuilt “IS A” SNOMED-CT hierarchy were also extracted and counted towards the corresponding target concept. For example, if the concept “juvenile myoclonic epilepsy (disorder)” was detected this would be counted towards the target term “epilepsy (disorder)”. Slight modifications to the grouping of extracted SNOMED-CT concepts were made and are as follows: SNOMED-CT does not provide a mapping from ASM drugs to their brand names, for example the brand Keppra(™) is not categorised under levetiracetam. To accommodate for free-text containing these alternative names, each of the ASM target concepts were enriched with brand names from the Electronic Medicines Compendium (EMC). The EMC is an actively maintained database containing approved and regulated prescribing and patient information for licensed medicines (Electronic Medicines Compendium, 2021). The full list of all target and number of children concepts are displayed in supplementary materials item 2A and 2B.

6.5.4 Data and Statistical Analyses

All data and statistical analyses were performed in python 3.8 using the python packages: medcat v0.4, sklearn v0.24, scipy v1.7, and pandas v1.3. Data visualisations were created using matplotlib v3.4, and plotly v5.1.

6.6 Results

The results for this chapter build upon the finding in the previous results chapter 3.

6.6.1 Demographic Analysis

Cogstack retrieved 36,855 documents of 9,860 unique patients who attended a neurology clinic between 2013-2020. Despite all documents containing the word “epilepsy”, MedCAT could accurately identify 4,011 patients who had a diagnosis of Epilepsy across 25,291 documents. The patients which were removed had no mention

of an epilepsy diagnosis within the 7-year period and their documents only contained phrases such as: “Seen in Epilepsy Clinic”, “Seen by Epilepsy Nurse”, “signed by Dr XXX Epilepsy consultant”. The proportion of females with epilepsy was 50.04%.

6.6.2 Data Extraction and Performance

After a subset of 400 documents were annotated by EB, PV, SE, CM, the performance of the MedCAT model was then evaluated against their annotations. This step was identical to the steps outlined in the previous chapter and therefore the results are the same. The full performance of annotations can be found in supplementary material items 3A and 3B.

As the aim of this project was to examine patterns of ASM utilisation, we excluded patients who did not undergo ASM intervention. The ascertainment of whether a patient was taking an ASM was an explicit mention that they were taking an ASM within their neurology clinic letters describing a clinic visit within the duration of this study. The total number of patients without any documentation of ASM within their neurology clinic letters were n=446 (11.12% of 4011). The number of patients with at least one documentation of an ASM in this study was n=3565 (88.89% of 4011).

6.6.3 ASM Prescribing Patterns

From all patients with a diagnosis of epilepsy the top 5 monotherapy and top 5 polytherapy ASM prescribing patterns are represented in Table 3. The number of patients who have tried these combinations, within the duration of the study, was n=2,215 (55.22% of 4011). Seizure freedom in this study is defined as when there is an explicit statement or variations of “seizure free” in the document. Mentions of “No seizures” or the absence of any seizure documentation was not treated as seizure free.

ASM prescribing pattern	Percentage of total people with epilepsy who have been on this regime	Seizure freedom rates whilst on the regime
Levetiracetam	17.3%	26.09%
Lamotrigine	13.4%	22.38%
Valproate	12.4%	22.38%
Carbamazepine	11.4%	32.60%
Levetiracetam + Lamotrigine	6.5%	18.70%
Levetiracetam + Carbamazepine	5.0%	26.52%
Lamotrigine + Valproate	4.7%	23.21%
Levetiracetam + Valproate	4.7%	24.55%
Topiramate	3.1%	28.57%
Levetiracetam + Lamotrigine + Valproate	2.4%	23.26%
Other ASM / combination	14.3%	26.17%

Table 3. In descending order of popularity. The top 5 monotherapy and top 5 polytherapy AED prescribing patterns and rates of seizure freedom.

We then examined the patterns of top 10 ASM patterns and described the proportion of drug switching between them (Figure 18). Here we show the patterns of ASM drug prescription patterns and show any switching that occurred anytime within the 7-year period: 34.5% switched once or more, 9.2% switched twice or more, 2.2% switched thrice or more, and 0.3% switched four or more times.

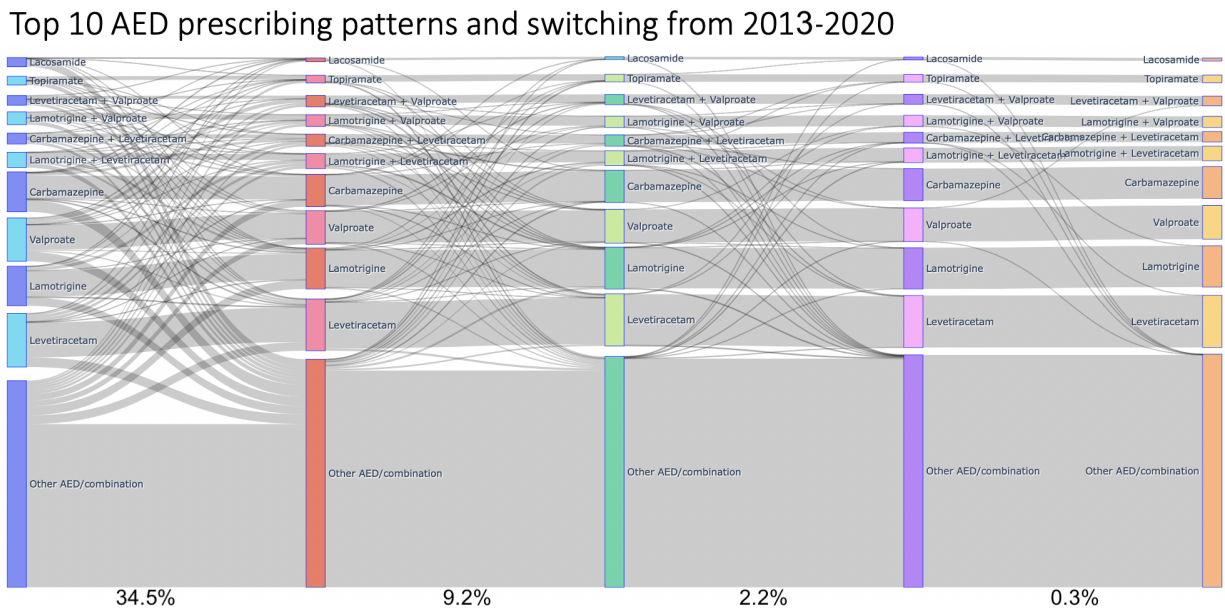


Figure 18. Top 10 ASM prescribing patterns and switching. The Sankey diagram represents the proportion of patients that have switched ASM medication within the timeframe of this study. Straight lines from one ASM to the identical adjacent label indicate the proportion of patients who no longer switched to another ASM/combination category. The bottom legend percentage represents the total proportion of patients that switched ASM/combination.

6.6.4 ASM Associated Symptoms

The frequency of common associated symptoms was explored. Figure 19 describes the common associated symptoms from the top 10 ASM prescribing patterns. The presence of anxiety and headaches were the most associated symptoms across all ASMs and combinations. Topiramate prescription is highly associated with headaches and nausea. Lamotrigine and its use in combinations is associated with rashes when compared to other ASMs and in combinations without.

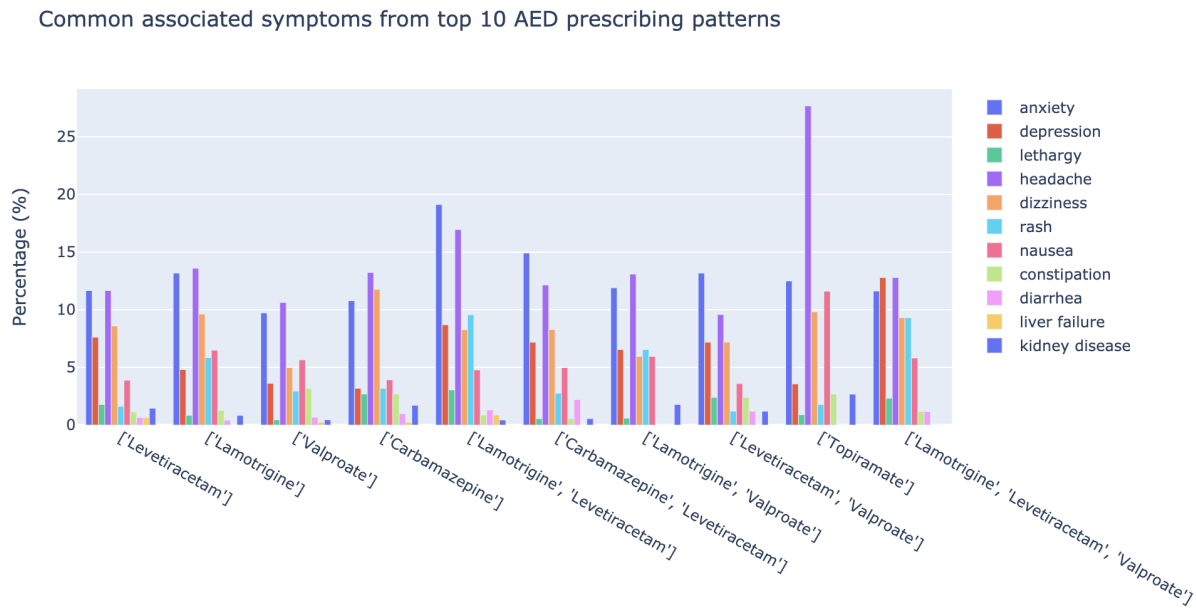


Figure 19. Percentage of common associated symptoms from the 10 top ASM prescribing patterns. Listed in descending order of prescription popularity from Left to Right. Note that the association does not imply forward causation of an adverse event, e.g. Headache associated with Topiramate might be a feature of Topiramate being used to treat migraines rather than an adverse event.

The overall comparison of associated symptoms between monotherapy and polytherapy is shown in figure 20. Nearly all associated symptoms increased in incidence during polytherapy when compared to monotherapy. Depression, rash and kidney disease exhibited the largest increase of approximately 2-fold. Only liver failure and lethargy did not exhibit any major increase.

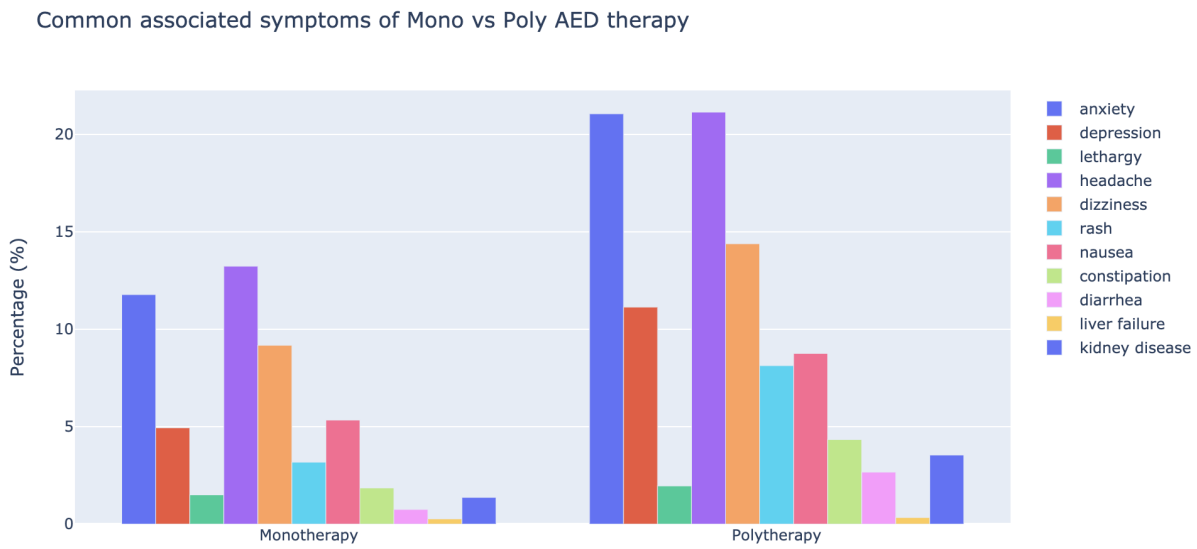


Figure 20. Percentage of common associated symptoms with ASM monotherapy and polytherapy

6.7 Discussion

The automatic extraction of epilepsy related concepts, ASM and other epilepsy variables relevant to monitoring treatment outcomes was feasible. In unstructured text, ASMs are unique in their nomenclature and do not often rely on the context for disambiguation. The word 'fit' would be an example of where further supervised training was required to help provide contextual examples for the model to 'learn' and link the correct SNOMED concept. "He was fit and well" and "The epilepsy patient had a fit" with the latter 'fit' example labelled to seizure would help the model disambiguate identical words with alternative contextual meanings. Consequently, the extensive coverage of SNOMED-CT combined with the SNOMED-CT UK drugs extension allowed for near out of the box accurate concept extraction with the CogStack-MedCAT pipeline. The only modification to the SNOMED-CT terminology structure which was made was the mapping of ASM primary active ingredient to commercial ASM brand names e.g. Tegretol, Tegretol XR, and Carbatrol are all medicinal products which contain carbamazepine. Although SNOMED-CT does support a mapping through the inbuilt relationship structure "has specific active ingredient" it is under-developed across the SNOMED-CT and UK drugs extension. Therefore, a local extension using the EMC database overcame this issue allowing the primary active ASM ingredient and any respective brand names to be standardised and extracted as a single concept.

6.7.1 Prescription Patterns

To date, the best source of evidence for the optimal choice of ASMs in new-onset epilepsy are the SANAD and SANAD II clinical trials, comparing the longer-term effects of ASM treatments (A. Marson et al., 2021a, 2021b; A. G. Marson, Al-Kharusi, et al., 2007a, 2007b). These studies are the largest open-label randomised controlled trials in hospital-based outpatient clinics in the UK that compare the effectiveness and cost effectiveness of monotherapy ASMs for new-onset epilepsy. The results have influenced UK national guidelines such as epilepsy treatment recommendations from the National Institute for Health and Care Excellence (NICE). The SANAD study is divided into two arms (Arm A and Arm B). The SANAD arm A concluded that lamotrigine was superior to all the other drugs for the treatment of focal epilepsy and

SANAD arm B concluded that valproate showed better efficacy than lamotrigine and better tolerability than topiramate for generalised epilepsy. The SANAD II trials were published after the time period of our studies and therefore its findings would have not influenced the prescription patterns in our study. The SANAD II trials determined that as first-line treatments for patients with focal epilepsy, lamotrigine was found to be superior in cost-effectiveness, and achieved lower rates of adverse reactions when compared to levetiracetam or zonisamide (A. Marson et al., 2021a). Similarly, levetiracetam when compared with valproate was found to be neither clinically effective nor cost-effective in patients with newly diagnosed idiopathic generalised or difficult to classify epilepsy (A. Marson et al., 2021b). Regarding these results, the SANAD authors consider lamotrigine to be the optimal treatment option in focal epilepsy and valproate the best in generalised epilepsy (although there are important limitations in using valproate in women of child-bearing potential). However, our findings may not necessarily reflect current medical practice of ASM prescriptions.

Prior to the findings of the SANAD trials, a 2004 national UK study examining ASM prescribing patterns found that the most common monotherapies across n=1,499 people with epilepsy were carbamazepine (37.4%), valproate (35.7%), phenytoin (29.4%), phenobarbitone or primidone (14.2%) and lamotrigine (10.3%) (Moran et al., 2004). Although the sample size for our study is nearly double (moran et al. n=1,499 to our study n=3,565) the national study's findings would be more generalisable as it samples a wider clinical population and demographic, yet would reflect former prescription patterns. Keeping that in mind, in comparison, we found an increase in the prescriptions of lamotrigine and levetiracetam and a decrease in carbamazepine, valproate, and phenytoin. At KCH from 2013 to 2020, the most popular ASMs were levetiracetam followed by lamotrigine. Despite carbamazepine, phenytoin, primidone, and valproate being amongst the earliest medications licensed for treating epileptic seizures, their long term use of have been associated with important adverse effects such as decreased bone mineral density that may lead to osteopenia, osteoporosis, and increased fracture risk (Feldkamp et al., 2012; Petty et al., 2005). Additionally these medications have potential adverse drug interactions, in the serum concentration of other ASMs, as well as other drug classes (anticoagulants, oral contraceptives, antidepressants, antipsychotics, antimicrobial drugs, antineoplastic drugs, and immunosuppressants) (Johannessen & Landmark, 2010). Therefore, it is

unsurprising that in place of carbamazepine, valproate and phenytoin; levetiracetam and lamotrigine have risen in popularity to become the most frequently used ASMs.

First-line ASMs guidelines may explain the popularity for several ASM prescription patterns. Epilepsy treatment recommendations from NICE state that: for treating newly diagnosed focal seizures, carbamazepine and lamotrigine are first-line options; whereas for newly diagnosed generalised tonic-clonic seizures, sodium valproate is the first-line treatment (except in female patients who are premenopausal) in these cases lamotrigine is the alternative choice, but may exacerbate myoclonic seizures if present (NICE, 2014, 2021b). Consequently, we would expect that these first-line ASMs to be the most popular especially for monotherapy and although carbamazepine, lamotrigine, and valproate were amongst the most popular ASMs, levetiracetam was the most popular. Although not recommended by NICE, levetiracetam is an increasingly popular first-line ASM with broad utility as it can be used for both focal and generalised epilepsies. In clinical practice this has numerous benefits especially for newly diagnosed epilepsy where ongoing investigations which have yet to determine the type of epilepsy and in circumstances where early attempts at unknown seizure type control are a necessity. In line with our study's finding, levetiracetam when compared to lamotrigine, has been shown to be well-tolerated and as effective in achieving seizure freedom (Rosenow et al., 2012). In elderly populations, levetiracetam has been shown to be superior to extended-release carbamazepine in terms of tolerability due to lower risk of adverse events, and absence of interactions with other drugs (Werhahn et al., 2015). Overall with levetiracetam's broad utility, comparable rates of adverse events and relatively lower potential for drug-drug interactions, the increasing use of levetiracetam in current clinical practice shouldn't be surprising.

If the first-line ASM are insufficient to control seizures or there are adverse side effects, ASM switching or ASM combinations may be required. We showed that amongst the top 5 monotherapy and polytherapy ASMs, 34.5% switched once or more, 9.2% switched twice or more, 2.2% switched thrice or more, and 0.3% switched four or more times. These rates of switching could be considered as a proxy measure of suboptimal ASM response. There is little to no evidence of what is the optimal ASM order or combinations. The majority of previous UK studies have looked at overall prescriptions

of ASMs in primary care without identifying their clinical population (Nicholas et al., 2012; Purcell et al., 2002; Roberts et al., 1998). This would have misrepresented the prescription patterns for epilepsy as ASM are commonly used for other conditions such as, psychiatric disorders and pain, and therefore these clinical groups would have been included within their study design (Amann et al., 2007; Chong & Smith, 2000).

6.7.2 Rates of Seizure Freedom

In terms of seizure freedom, we found that carbamazepine monotherapy has the highest rate of seizure freedom, followed by topiramate, and then levetiracetam. Seizure freedom rates were observed to be lower for ASM polytherapies which likely reflects the refractory nature of the seizures in which they are required to be tried in the first place. From the top 5 combinations levetiracetam + carbamazepine, lamotrigine + valproate, levetiracetam + valproate, and levetiracetam + lamotrigine + valproate, displayed equivalent seizure freedom rates and are comparable to other studies findings of 20–30% of patients achieving seizure freedom (Mohanraj & Brodie, 2005; Peltola et al., 2008). Whilst levetiracetam + lamotrigine although the most common combination displayed the lowest seizure freedom rates (18.7%).

These findings of rates of seizure freedom are likely to vary according to what is generally considered to be seizure free (no seizures within a 2-year period). In our study, a patient was attributed to being seizure free if there was an explicit documentation of seizure freedom alongside no concurrent mentions of recent seizures. The pattern of how clinicians use the expression when a patient is seizure free would influence the results and the extent of this impact is unknown. Overestimates would occur if clinicians loosely mentioned that the patient had seizure free periods and did not report any seizures alongside. Whilst underestimates would occur if explicit seizure freedom phrases were under-used. Phrases such as “seizures under control, no change in medication” may be meant by the author as seizure freedom but were not interpreted as such. In our study, seizure control was not interpreted as seizure free, and the absence of mentions of seizure occurrence was also not treated as evidence of seizure absence. Additionally, our study is at risk of survivorship bias as patients are less likely to attend if seizure-free and therefore their seizure free state was not documented.

When examining ASM efficacy, an alternative metric to seizure freedom is seizure frequency. However, the greatest limitation to assessing seizure frequency from EPR data is that it cannot be extracted reliably from the record using currently available techniques. This is due to the irregularity and variability in how seizure frequency is documented within and across EPR documents. Assessing seizure frequency accuracy is crucial and can serve as a surrogate of seizure control efficacy. The development of a standardised reporting terminology which has the capacity to characterise all different seizure frequencies is crucial to this. Clinicians would then be able to either directly enter this information in the EPR from a structured drop-down list or a post processing step (searching for rate terms surrounding the seizure concept). Alternatively, NLP relational extraction techniques (linking seizure concepts to rates based on linguistic patterns) (Wang et al., 2022) would be required to accurately standardise, extract, and analyse this information.

6.7.3 Idiosyncratic Side Effects/Associated Symptoms

We know that there are many individual side effects and toxicities of ASMs associated at an individual drug level, class of drugs level and also importantly multiple drug-drug interactions (Patsalos et al., 2002). The presence of ASM adverse reactions may influence the choice of a specific ASM or combination. A drug that is only modestly efficacious but has a favourable safety profile may look better than a drug that is more efficacious but produces clinically meaningful adverse events. Therefore, a drug's retention rate is mainly determined by its side effect profile (Bootsma et al., 2009).

Extracting concepts from electronic health records can only shed light on the correlation between diseases and or medications/combinations. For example, anxiety may not necessarily be caused by an adverse event to an ASM, but could be attributed to their risk of seizures or other external life events. This is why the term ‘associated’ is used rather than adverse reaction or side effect. Straightforward measurement of associations between different diseases, medications, and side effects has been referred to as an “empirical statistical phenomenon that has no meaning in itself” (Rutter, 1997). However, measurement of associations is an initial step in a process which in turn is followed by investigations into why certain conditions have an

increased risk to co-occur. Nevertheless, our study does reflect well known side effect phenomena such as the association of rash with lamotrigine (Brodie et al., 1995; Messenheimer et al., 1994); or the increased risk of kidney stone formation during topiramate use (Dell'Orto et al., 2014; Vega et al., 2007).

In the other direction of effect, the presence of comorbidity may result in one ASM being selected over another due to its beneficial effects on conditions beyond seizure control. This means that high rates of an associated symptom may mean that the drug is prescribed for the associated symptoms. An example of this is our finding of a high association between headaches and topiramate treatment. It is well known that Topiramate is prescribed for migraine prevention (Diener et al., 2004; Silberstein et al., 2007) hence the association between topiramate and headache is likely driven by this factor, rather than headache being a common topiramate side-effect. Overall the rationale of ASM choice and ASM side effect phenomena are intertwined in our findings.

Nearly all associated symptoms increased in incidence during polytherapy when compared to monotherapy. Depression, rash, and kidney disease exhibited the largest increase of approximately 2-fold. Only liver failure and lethargy did not exhibit any major increase. A limitation of this study was that ASM dosages were not looked at in this study. ASM dosages were not explored and this could have been a major factor in the presence of associated symptoms. Future studies should examine routinely collected ASM blood serum concentration levels. Although not all patients are required to have ASM levels taken, this information is usually structured and could be used to measure ASM compliance, validate the performance of the NLP model and even provide further insights into seizure freedom.

The most popular ASM combination was levetiracetam+lamotrigine which also displayed the lowest rate of seizure freedom. Despite evidence that the pharmacokinetic profile of levetiracetam is not influenced by lamotrigine and that newer AEDs are generally considered effective, and many have favourable safety profiles (Brodie & Kwan, 2012; Gidal et al., 2005; Golyala & Kwan, 2017). The findings of this study suggest that there appears to be a synergistic increase in associated symptoms when compared to when they are taken separately.

In recent times, the most significant ASMs side effects findings has been the identification of the link between sodium valproate and congenital abnormalities (Campbell et al., 2013; Vajda et al., 2013). This has quickly led to therapeutic policy changes that state; sodium valproate should not be used during pregnancy and in women of childbearing potential. These observational studies uncovered the effect through building evidence constructed from self-reports by patients, as well as reporting by healthcare professionals. Although this study did not look into pregnancy and congenital abnormalities, there is no reason why these disorders and others couldn't be added to the pipeline for further insights into trends of medication use and health outcomes. In the future, active automated extraction of side-effects and associated symptoms from interoperable EPR systems will have the capability of rapid identification of trends in adverse events, increasing patient safety.

In conclusion there are multiple reasons why a patient may switch their ASM or combination. The primary reason is for seizure control and the second to minimise side effects from these drugs. There is evidence that ASM prescription patterns vary in accordance to age and gender differences (Landmark et al., 2011). In addition, different ASMs are effective against different seizure types and different epilepsy syndromes (Thijs et al., 2019). However, drugs that have been shown to be effective against focal seizures may be relatively ineffective against some generalised seizures. Conversely, those effective against generalised seizures may be somewhat effective against seizures with focal onset. During the diagnostic process and evolution in the course of a patient's epilepsy disorder, improved clarity about seizure types and comorbidities will influence any changes to ASM choice. Prescription patterns and their interactions with demographics, comorbidities, seizure and epilepsy types could be subsequently analysed in future studies being primarily limited by the degrees of freedom of comparisons.

6.8 Conclusions

A wealth of information regarding ASM efficacy lies within electronic patient records (EPRs), yet efficient data extraction has remained a critical barrier to closing this knowledge gap. An automated information extraction technique applied to electronic

patient records, particularly free-text, can actively be used to explore insights beyond direct patient care and into treatment strategies, and associated symptoms.

The present study highlights the patterns of ASM use and lists their rates of associated symptoms. We found that levetiracetam was the most common ASM monotherapy and levetiracetam+lamotrigine the most common combination. The ASM combination of levetiracetam+lamotrigine displayed the lowest rate of seizure freedom as well as a synergetic increase in rate of associated symptoms when compared to patients taking these drugs individually.

Overall, determining the optimal choice in ASM, combination and switching order is challenging. Demographic variables, such as age and gender, comorbidities, and comedication play a role in how clinicians choose which ASM to trial, exacerbating non-random assignment to treatment groups. Underlying epilepsy aetiology, baseline epilepsy severity, and age of onset also correlate with refractoriness to ASM treatment. Holistic big data approaches which monitor real-world clinical practice have the potential to continuously monitor current therapeutic practises, outcomes, and any potential changes. These methodologies should be considered as supplementary to clinical trial evidence to inform best clinical practice. Together these can improve the management of epilepsies through evidence-based decisions.

Chapter 7: Generalisability of These Tools Beyond Epilepsy and Across Hospitals

During the course of this thesis there was a global pandemic caused by COVID-19. An opportunity to demonstrate that the CogStack/MedCAT pipeline can generalise to and benefit other medical specialities was presented. A full list of all COVID related works are listed in the thesis disseminations.

This chapter contains formatting adjustments which do not follow the style of the rest of the thesis. It is divided into two sections: The first section demonstrates the applicability of information extraction and the MedCAT pipeline to other areas of medicine. The second section demonstrates the external validity of an NLP model through monitoring the performance of information extraction models and its validation at other large NHS hospital trusts.

7.1 Pre-existing Cardiovascular Disease Rather than Cardiovascular Risk Factors Drives Mortality in COVID-19

7.1.1 Authors

Kevin O'Gallagher*, Anthony Shek*, Dan Bean, Alexandros Papachristidis, James T H Teo, Richard J. B. Dobson, Ajay M. Shah, Rosita Zakeri.
*These authors contributed equally

This work was published in the journal: BMC cardiovascular disorder
DOI: 10.1186/s12872-021-02137-9

O'Gallagher *et al. BMC Cardiovasc Disord (2021) 21:327*
https://doi.org/10.1186/s12872-021-02137-9

BMC Cardiovascular Disorders

RESEARCH ARTICLE Open Access

Pre-existing cardiovascular disease rather than cardiovascular risk factors drives mortality in COVID-19

Kevin O'Gallagher[1†], Anthony Shek[2†], Daniel M. Bean[2], Rebecca Bendayan[2,3], Alexandros Papachristidis[4], James T. H. Teo[4], Richard J. B. Dobson[2,5], Ajay M. Shah[1,4,6*†] and Rosita Zakeri[1,6*†]

Abstract

Background: The relative association between cardiovascular (CV) risk factors, such as diabetes and hypertension, established CV disease (CVD), and susceptibility to CV complications or mortality in COVID-19 remains unclear.

Methods: We conducted a cohort study of consecutive adults hospitalised for severe COVID-19 between 1st March and 30th June 2020. Pre-existing CVD, CV risk factors and associations with mortality and CV complications were ascertained.

Results: Among 1721 patients (median age 71 years, 57% male), 349 (20.3%) had pre-existing CVD (CVD), 888 (51.6%) had CV risk factors without CVD (RF-CVD), 484 (28.1%) had neither. Patients with CVD were older with a higher burden of non-CV comorbidities. During follow-up, 438 (25.5%) patients died: 37% with CVD, 25.7% with RF-CVD and 16.5% with neither. CVD was independently associated with in-hospital mortality among patients < 70 years of age (adjusted HR 2.43 [95% CI 1.16–5.07]), but not in those ≥ 70 years (aHR 1.14 [95% CI 0.77–1.69]). RF-CVD were not independently associated with mortality in either age group (< 70 y aHR 1.21 [95% CI 0.72–2.01], ≥ 70 y aHR 1.07 [95% CI 0.76–1.52]). Most CV complications occurred in patients with CVD (66%) versus RF-CVD (17%) or neither (11%; $p < 0.001$). 213 [12.4%] patients developed venous thromboembolism (VTE). CVD was not an independent predictor of VTE.

Conclusions: In patients hospitalised with COVID-19, pre-existing established CVD appears to be a more important contributor to mortality than CV risk factors in the absence of CVD. CVD-related hazard may be mediated, in part, by new CV complications. Optimal care and vigilance for destabilised CVD are essential in this patient group. *Trial registration* n/a.

Keywords: COVID-19, Cardiovascular disease, Cardiovascular risk factors, Hypertension, Diabetes

Background

Cardiovascular (CV) risk factors such as hypertension and diabetes, and chronic CV diseases (CVD), including ischaemic heart disease and heart failure, are highly prevalent among patients admitted to hospital with severe novel coronavirus disease 2019 (COVID-19) [1–5]. In population-based studies, diabetes and chronic CVD, but not hypertension, have been associated with higher mortality [6, 7]. At present, patients with either established CVD or CV risk factors are considered to be vulnerable

*Correspondence: ajay.shah@kcl.ac.uk; rosita.zakeri@kcl.ac.uk
†Kevin O'Gallagher and Anthony Shek have contributed equally to this work.
†Ajay M. Shah and Rosita Zakeri have jointly led and contributed to the study.
[6] School of Cardiovascular Medicine and Sciences, James Black Centre, King's College London, 125 Coldharbour Lane, London SE5 9NU, UK
Full list of author information is available at the end of the article
All authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

individuals [8]. However, it remains unclear whether an increased susceptibility to severe COVID-19 in patients with CV risk factors is driven by co-existent CVD, or whether patients with CV risk factors without established CVD have a similarly severe course.

The relationship between CVD and COVID-19 may also be bidirectional. SARS-CoV-2 is reported to directly infect the endothelium and possibly the heart [9, 10], which could precipitate CV complications. Isolated case reports of fulminant myocarditis or pericarditis have been attributed to COVID-19 [11–13], although the incidence and mechanism of such complications is debated. Furthermore, while patients with pre-existing CVD may be at increased risk of CV complications [14, 15], it is not clear the extent to which these represent recurrent or decompensated CVD rather than de novo complications, nor whether the risk also applies to patients with CV risk factors.

To address these questions, we evaluated outcomes associated with pre-existing CVD and CV risk factors, in a large multi-ethnic cohort of patients hospitalised for severe COVID-19. Our aims were to determine (a) the relative risks of in-hospital mortality and CV complications for individuals with COVID-19 and pre-existing CVD versus CV risk factors without established CVD, and (b) factors associated with the occurrence of major CV complications in patients with COVID-19.

Methods

Approvals

This study was conducted under London South East Research Ethics Committee approval (reference 18/LO/2048) granted to the King's Electronic Records Research Interface (KERRI); COVID-19 work was reviewed with expert patient input on a virtual committee with Caldicott Guardian oversight.

Study design

We conducted a cohort study of consecutive adult patients (age > 18 y) admitted with COVID-19 to King's College Hospital NHS Foundation Trust (comprising King's College Hospital and Princess Royal University Hospital), between 1st March and 30th June 2020. All patients had a positive RT-PCR antigen test for SARS-CoV-2. Only patients admitted to hospital for ≥ 24 h were included. A subset of this cohort has been reported previously [3, 16].

Data sources and processing

Structured and unstructured data were extracted from the electronic health record (EHR) using previously described natural language processing (NLP) informatics tools belonging to the CogStack ecosystem [17], DrugPipeline [18], MedCAT [19], and MedCATtrainer [20]. Clinician case review was used for additional validation (Additional file 1: Methods).

Exposures and outcomes

CV risk factors were defined as a recorded clinical diagnosis of hypertension, diabetes mellitus, or self-reported smoking status, in the absence of documented CVD. Diabetes mellitus was defined by a clinician diagnosis documented in the electronic health record and extracted based on relevant SNOMED CT UK extension and children terms (S-44054006, S-73211009, S-46635009) encompassing both type 1 and type 2 diabetes mellitus, as previously described [19] and validated [16]. Pre-existing established CVD was defined as a previous record of ≥ 1 of the following diagnoses: myocardial infarction (MI), heart failure, myocarditis, pericarditis, endocarditis, atrial or ventricular arrhythmia, and severe valvular heart disease [21]. Additional details are provided in Additional file 1: Methods. CV risk factors and pre-existing CVD were categorised as present if they had been recorded in the EHR at any time up to the day of admission (or including the day of admission, when recorded as a pre-existing condition). Data were also collected for age, sex, ethnicity, body mass index (BMI), non-CV comorbidities (asthma, chronic obstructive pulmonary disease [COPD], chronic kidney disease [CKD]), previous venous thromboembolism [VTE] comprising deep vein thrombosis [DVT] or pulmonary embolism [PE]), CV drug therapy (Additional file 1: Methods) and clinical examination and routinely collected blood results on admission. High sensitivity cardiac troponin T (hs-cTnT) plasma levels were defined as normal when below the 99th percentile of normal values, i.e., 14 ng/L.

The primary outcome was in-hospital mortality, with cause of death ascertained from death certification. Secondary outcomes included any CV complication related to COVID-19 and incident VTE. A CV complication was defined as a new CVD diagnosis or decompensation of pre-existing CVD recorded in the EHR on presentation or at any time during admission. CV complications were based on clinician diagnoses incorporating all available clinical information, including where relevant, echocardiography and coronary angiography (additional details in Additional file 1: Methods). Hospital admission date was used as the start of follow-up. Outcomes were ascertained through to death, discharge, or 31st July 2020, whichever was earlier.

Statistical analyses

Patient data are reported as frequency (%), mean (SD) or median (IQR), as appropriate. Patient characteristics were compared across 3 groups: patients with

pre-existing established CVD (CVD), CV risk factors without CVD (RF-CVD), and no CVD or CV risk factors, using the Chi-squared goodness of fit or Fisher's exact test (categorical variables), one-way analysis of variance (continuous variables) or Kruskal–Wallis/Wilcoxon rank-sum tests (for non-normally distributed data). Bonferroni correction was used for individual comparisons. Missing blood biomarkers (< 25% missing data) were imputed using the multiple imputation approach by chain equations [22].

Cumulative incidence plots displaying the probability of in-hospital mortality and discharge were constructed based on a competing risks analysis. To evaluate the association between patient group and mortality, we used Cox proportional hazards regression, with admission date as the start of follow-up and in-hospital mortality as the dependent variable. Unadjusted, demographic adjusted (age, sex, ethnicity), and fully adjusted models (including non-CV comorbidities and medications on admission) were performed. Age was modelled as a categorical variable to allow for potentially non-linear association (< 40, 40–49, 50–59, 60–69, 70–79, 80+ years). Comorbidities were modelled as binary variables. The reference group comprised patients without pre-existing CVD or CV risk factors. The proportional hazard assumption was examined graphically and using formal tests, as described by Grambsch [23]; no major deviations from this assumption were observed.

To investigate the association between CV complications and prognosis we performed logistic regression models with patient groups stratified by the presence or absence of CV complications as an independent categorical variable. For secondary outcomes of CV complications or VTE, logistic regression models were constructed with (1) patient group and (2) individual CV risk factors or CVDs as binary predictor variables. Unadjusted, demographic-adjusted, and fully adjusted regression models were performed as above. When individual CVDs were examined, myocarditis and pericarditis were excluded, due to their low prevalence. Information regarding hyperlipidaemia diagnoses were incomplete. A sensitivity analysis was performed where all patients who were prescribed statin therapy in the absence of diagnosed CVD were reclassified as RF-CVD. Furthermore, as BMI was missing in > 30% of patients, primary analyses were performed without adjustment for BMI. However, sensitivity analyses were performed restricted to patients with BMI data available including (1) adjustment for BMI as a continuous variable (fully adjusted model), (2) adjustment for obesity as a categorical variable (defined as BMI ≥ 30 kg/m^2), and (3) reclassifying patients with obesity without diagnosed CVD as RF-CVD. An additional sensitivity analysis was performed among individuals with outcome data available i.e. those who were discharged or died (excluding current in-patients). Analyses were performed using STATA/IC (v16.1; StataCorp LLC, TX).

Results

Study population

Between 1st March and 30th June 2020, 1,721 patients were admitted with COVID-19 (median age 71 years [IQR 56–83], 56.6% male). Of these, 349 (20.3%) had CVD, 888 (51.6%) had RF-CVD, and 484 (28.1%) had neither. Patients with CVD were older than patients with RF-CVD or neither but had a similar distribution of sex (Table 1).

CVD was more prevalent with increasing age, while RF-CVD was most common between 50–70 years (Additional file 3: Fig. S1A). Individuals from non-White ethnic groups had a higher prevalence of RF-CVD whereas CVD was more prevalent in the White group (Additional file 3: Fig. S1B). The most frequent CVD diagnoses were arrhythmia (86.2% atrial fibrillation), heart failure, and previous MI, respectively (Table 1). 119 (34.1%) patients with CVD had more than one CVD diagnosis. Rates of non-cardiovascular comorbidities were highest in patients with CVD, followed by RF-CVD (Table 1).

On admission, 83% of patients with hypertension were taking an antihypertensive agent and 73% of patients with atrial fibrillation were on oral anticoagulation. Rates of ACEI or ARB and betablocker use in heart failure patients were 47% and 62% respectively. In patients with a previous MI, rates of antiplatelet, beta-blocker and statin use were 68%, 64% and 65% respectively.

Clinical presentation

Physiological parameters and blood biomarkers are displayed in Additional file 2: Table S1. There were few clinically significant differences in physiological observations between groups, with the exception of a higher mean systolic blood pressure in patients with RF-CVD. Among blood biomarkers, C-reactive protein values were highest in patients with RF-CVD, but similar between patients with CVD and the group with neither CVD nor risk factors (Additional file 2: Table S1). Renal function was progressively worse across groups, with the lowest eGFR in patients with CVD.

Overall, 742 (43.1%) patients had at least one hs-cTnT measurement. Among patients with at least one hs-cTnT measurement, elevated values (> 14 ng/L) were observed in 133/147 (90.5%) patients with CVD, 311/409 (76.0%) patients with RF-CVD and 123/186 (66.1%) patients with no CVD or CV risk factors.

Table 1 Patient characteristics

	Total N = 1721	**No CVD or CV risk factors N = 484 (28.1%)**	**RF-CVD N = 888 (51.6%)**	**CVD N = 349 (20.3%)**	***p* value**
Demographics					
Age, y	71 (56–83)	58 (44–75)	71 (59–82)	81 (71–88)	< 0.001
Age group, n (%)					< 0.001
< 40	132 (7.7)	97 (20.0)	31 (3.5)	4 (1.2)	
40–49	127 (7.4)	64 (13.2)	55 (6.2)	8 (2.3)	
50–59	266 (15.5)	93 (19.2)	149 (16.8)	24 (6.9)	
60–69	303 (17.6)	73 (15.1)	187 (21.1)	43 (12.3)	
70–79	316 (18.4)	56 (11.6)	173 (19.5)	87 (24.9)	
80+	577 (33.5)	101 (20.9)	293 (33.0)	183 (52.4)	
Male sex	974 (56.6)	259 (53.5)	522 (58.8)	193 (55.3)	0.146
BMI, kg/m^2	26.2 (22.5–31.1)	25.4 (21.8–30.6)	26.9 (23.1–31.8)	25.1 (22.0–29.0)	< 0.001
*BMI category**					< 0.001
Underweight	74 (4.3)	23 (4.8)	29 (3.3)	22 (6.3)	
Normal weight	349 (20.3)	103 (21.3)	152 (17.1)	94 (26.9)	
Overweight	291 (16.9)	71 (14.7)	155 (17.5)	65 (18.6)	
Obese	317 (18.4)	83 (17.2)	177 (19.9)	57 (16.3)	
Missing	690 (40.1)	204 (42.2)	375 (42.2)	111 (31.8)	
Ethnicity					< 0.001
White	845 (49.1)	238 (49.2)	382 (43.0)	225 (64.5)	
Black	434 (25.2)	85 (17.6)	280 (31.5)	69 (19.8)	
Asian	96 (5.6)	29 (6.0)	55 (6.2)	12 (3.4)	
Mixed/other	121 (7.0)	37 (7.6)	64 (7.2)	20 (5.7)	
Missing	225 (13.1)	95 (19.6)	107 (12.1)	23 (6.6)	
Comorbidities					
Cardiovascular risk factors					
Hypertension	963 (56.0)	–	689 (77.6)	274 (78.5)	0.726
Diabetes	601 (34.9)	–	440 (49.6)	161 (46.1)	1.000
Type 1	6 (0.4)	–	5 (0.6)	1 (0.3)	0.305
Type 2	595 (34.6)	–	435 (49.0)	160 (45.9)	0.320
Ever smoker	432 (25.1)	–	314 (35.4)	118 (33.8)	0.607
Current smoker	114 (6.6)	–	85 (9.6)	29 (8.3)	0.490
Ex-smoker	318 (18.5)	–	229 (25.8)	89 (25.5)	0.917
Cardiovascular diseases					
Previous myocardial infarction	107 (6.2)	–	–	107 (30.7)	–
Chronic heart failure	133 (7.7)	–	–	133 (38.1)	–
Previous myocarditis	4 (0.2)	–	–	4 (1.2)	–
Previous pericarditis	3 (0.2)	–	–	3 (0.9)	–
Arrhythmia	218 (12.7)	–	–	218 (62.5)	–
Atrial fibrillation	188 (10.9)	–	–	188 (53.9)	–
Valvular heart disease**	15 (0.9)	–	–	15 (4.3)	–
Previous endocarditis	16 (0.9)	–	–	16 (4.6)	–
Non-cardiac comorbidities					
Asthma	148 (8.6)	16 (3.3)	74 (8.3)	58 (16.6)	< 0.001
COPD	129 (7.5)	5 (1.0)	53 (6.0)	71 (20.3)	< 0.001
Chronic kidney disease	165 (9.6)	6 (1.2)	71 (8.0)	88 (25.2)	< 0.001
Previous pulmonary embolism	90 (5.2)	8 (1.7)	27 (3.0)	55 (15.8)	< 0.001
Previous deep vein thrombosis	110 (6.4)	12 (2.5)	37 (4.2)	61 (17.5)	< 0.001

Table 1 (continued)

	Total N = 1721	No CVD or CV risk factors N = 484 (28.1%)	RF-CVD N = 888 (51.6%)	CVD N = 349 (20.3%)	*p* value
Medication					
ACEI/ARB	528 (31.4)	29 (6.5)	353 (39.9)	146 (41.8)	< 0.001
Aldosterone antagonist	66 (3.9)	11 (2.5)	24 (2.7)	31 (8.9)	< 0.001
Beta-blocker	430 (25.6)	43 (9.6)	197 (22.3)	190 (54.4)	< 0.001
Calcium-channel blocker	458 (27.3)	26 (5.8)	345 (39.0)	87 (24.9)	< 0.001
Loop diuretic	245 (14.6)	20 (4.5)	93 (10.5)	132 (37.8)	< 0.001
Statin	678 (40.3)	64 (14.3)	420 (47.5)	194 (55.6)	< 0.001
Anticoagulant	325 (19.3)	53 (11.8)	108 (12.2)	164 (47.0)	< 0.001
Antiplatelet agent	395 (23.5)	47 (10.5)	222 (25.1)	126 (36.1)	< 0.001
Metformin	299 (17.4)	–	240 (27.0)	59 (16.9)	–
Sulphonylurea	128 (7.4)	–	104 (11.7)	24 (6.9)	–
Repaglinide	–	–	–	–	–
SGLT2 inhibitor	23 (1.3)	–	17 (1.9)	6 (1.7)	–
DPP4 inhibitor	134 (7.8)	–	88 (9.9)	46 (13.2)	–
Thiazolidinedione	4 (0.2)	–	3 (0.3)	1 (0.3)	–
GLP1 receptor agonist	14 (0.8)	–	8 (0.9)	6 (1.7)	–
Insulin	180 (10.5)	–	128 (14.4)	52 (14.9)	–
COVID-19 investigational therapies					
Hydroxychloroquine	15 (0.9)	6 (1.2)	7 (0.8)	2 (0.6)	0.553
Dexamethasone/Prednisolone	204 (11.9)	50 (10.3)	103 (11.6)	51 (14.6)	0.159
Remdesivir	2 (0.1)	1 (0.2)	–	1 (0.3)	–
Colchicine	16 (0.9)	2 (0.4)	7 (0.8)	7 (2.0)	0.049
Tocilizumab	2 (0.1)	–	–	2 (0.6)	–
Azithromycin	26 (1.5)	13 (2.7)	8 (0.9)	5 (1.4)	0.035

Data represent n (%) or median (IQR)

p values refer to comparisons across 3 groups (except for cardiovascular risk factors, where comparisons are between 2 groups: CVD vs. RF-CVD)

ACEI angiotensin converting enzyme inhibitors, *ARB* angiotensin receptor blockers, *BMI* body mass index, *COPD* chronic obstructive pulmonary disease, *CV* cardiovascular, *CVD* cardiovascular disease, *DPP4* dipeptidyl peptidase-4, *GLP1* glucagon-like peptide 1, *RF-CVD* cardiovascular risk factors without established CVD, *SGLT2* sodium glucose co-transporter-2

*BMI categories classified as: underweight (< 18.5 kg/m^2), normal (18.5–24.9 kg/m^2), overweight (25–29.9 kg/m^2), and obese (≥ 30 kg/m^2)

**Severe degree of valvular heart disease (ESC guidelines [21])

In-hospital mortality

In-hospital outcomes are displayed in Table 2. Overall, 438 (25.5%) patients died and 1246 (72.4%) were discharged alive. 37 (2.1%) patients were in hospital at study close. The median length of hospitalisation for patients discharged was 9 (IQR 4–17) days and was longer for patients with CVD than those without (11 [IQR 5–19] vs. 7 [IQR 3–16] days, $p < 0.001$). Among patients who died, finalised death certificates were available in 382 patients. COVID-19 related pneumonia or acute respiratory distress syndrome was reported as the direct cause of death in 302 (79.1%) patients and as an indirect cause of death in 68 (17.8%) patients due to complications associated with COVID-19. This included 20 patients who died due to a CV cause: stroke (n = 8), massive PE (n = 3), decompensated heart failure (n = 4), myocardial infarction (n = 4) and acute limb ischaemia (n = 1). Only 12 deaths (3.1%) deaths were not attributed to COVID-19, e.g., malignancy or advanced dementia.

In-hospital mortality was greatest among patients with CVD (37.3%), intermediate in patients with RF-CVD (25.7%), and lowest among patients with neither (16.5%). Figure 1 displays cumulative incidence plots of the probability of in-hospital death or discharge over time for each group. For the overall cohort, there was a positive association between CVD and in-hospital mortality in unadjusted (HR 2.17 [95% CI 1.64–2.87], $p < 0.001$), and demographic-adjusted models (adjusted HR 1.45 [95% CI 1.09–1.94], $p = 0.012$). In fully adjusted models additionally accounting for non-CV comorbidities and baseline medications, there was a positive trend (aHR 1.36 [95% CI 0.97–1.92], $p = 0.076$).

Table 2 Complications and in-hospital outcomes of patients with COVID-19

	Total N = 1721	No CVD or CV risk factors N = 484 (28.1%)	RF-CVD N = 888 (51.6%)	CVD N = 349 (20.3%)	*p* value
Complications					
Cardiac					
Acute myocardial infarction	68 (4.0)	4 (0.8)	21 (2.4)	43 (12.3)	< 0.001
Acute heart failure	151 (8.8)	12 (2.5)	43 (4.8)	96 (27.5)	< 0.001
Myocarditis	12 (0.7)	0	9 (1.0)	3 (0.9)	0.090
Pericarditis	2 (0.1)	1 (0.2)	1 (0.1)	0	0.688
Arrhythmia*	314 (18.3)	40 (8.3)	99 (11.2)	175 (50.1)	< 0.001
Atrial fibrillation	266 (15.5)	28 (5.8)	74 (8.3)	164 (47.0)	< 0.001
Number of cardiac complications**					< 0.001
0	1,290 (75.0)	433 (89.5)	738 (83.1)	119 (34.1)	
1	325 (18.9)	45 (9.3)	129 (14.5)	151 (43.3)	
2+	106 (6.2)	6 (1.2)	21 (2.4)	79 (22.6)	
Venous thromboembolism					
Pulmonary embolism	151 (8.8)	41 (8.5)	66 (7.4)	44 (12.6)	0.015
Deep vein thrombosis	98 (5.7)	21 (4.3)	43 (4.8)	34 (9.7)	0.001
Extra-cardiac					
Acute kidney injury***	266 (15.5)	30 (6.2)	164 (18.5)	72 (20.6)	< 0.001
ARDS	77 (4.5)	19 (3.9)	46 (5.2)	12 (3.4)	0.324
Mechanical ventilation	92 (34.9)	36 (43.9)	41 (27.2)	15 (48.4)	0.009
Outcomes					
Died in hospital	438 (25.5)	80 (16.5)	228 (25.7)	130 (37.3)	< 0.001
ICU admission	226 (13.1)	75 (15.3)	127 (14.3)	24 (6.9)	< 0.001
Death or ICU admission	587 (34.1)	133 (27.4)	311 (35.0)	145 (41.4)	< 0.001
Discharged from hospital alive	1,246 (72.4)	393 (81.2)	639 (72.0)	214 (61.3)	< 0.001
Hospital length of stay[$], days	9 (4–17)	7 (3–16)	8 (4–18)	11 (5–19)	< 0.001

Data presented as n (%) or median (IQR). Table includes all in-hospital diagnoses during the admission (including new and recurrent diagnoses)

*Any physician-identified cardiac arrhythmia

**Number of physician-diagnosed CV complications from the following: acute myocardial infarction, heart failure, myocarditis, pericarditis, arrhythmia including AF, and endocarditis

***Acute kidney injury was defined according to the Kidney Disease: Improving Global Outcomes definition[39]

[$] Among patients discharged[40–43]

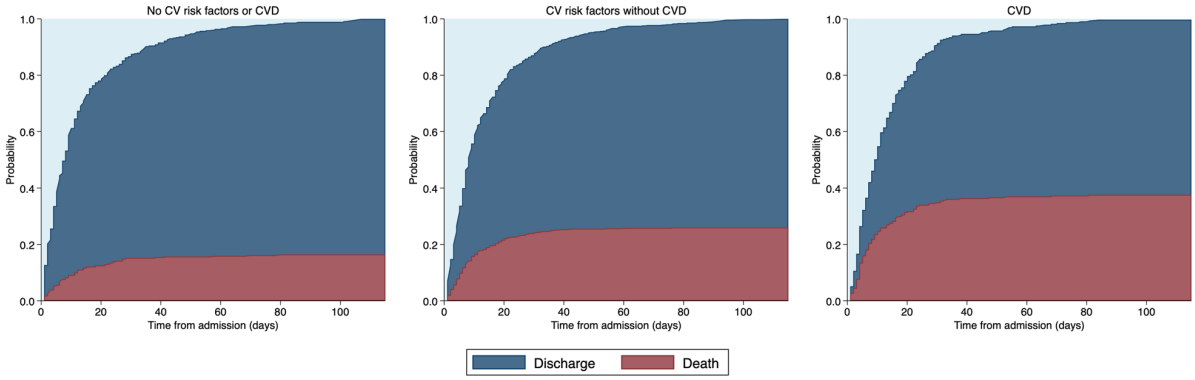


Fig. 1 Cumulative incidence plots displaying the probability of in-hospital death and discharge over time. The light blue region represents the probability of being alive and still in hospital at the time of study close. CV denotes cardiovascular; CVD, cardiovascular disease

This effect was principally driven by a prognostic association in patients under 70 years of age (Fig. 2A), whereas the effect of CVD was smaller and non-statistically significant among patients aged 70 years and older (Fig. 2B). RF-CVD conferred an increased risk of mortality for the overall cohort in unadjusted analyses (HR 1.51 [95% CI 1.17–1.95], $p = 0.002$), but not in demographic-adjusted (aHR 1.17 [95% CI 0.90–1.53], $p = 0.233$) or fully adjusted models (aHR 1.13 [95% CI 0.85–1.51], $p = 0.388$). RF-CVD were not associated with mortality in patients older or younger than 70 years of age (Fig. 2A, B).

The main findings were unchanged in a sensitivity analysis with patients prescribed statin therapy without diagnosed CVD reclassified as RF-CVD (n = 64 patients reclassified, Additional file 3: Fig. S2A, B). BMI data were available in 1031 patients (60% total cohort). Sensitivity analyses in this subset adjusting for BMI as a continuous

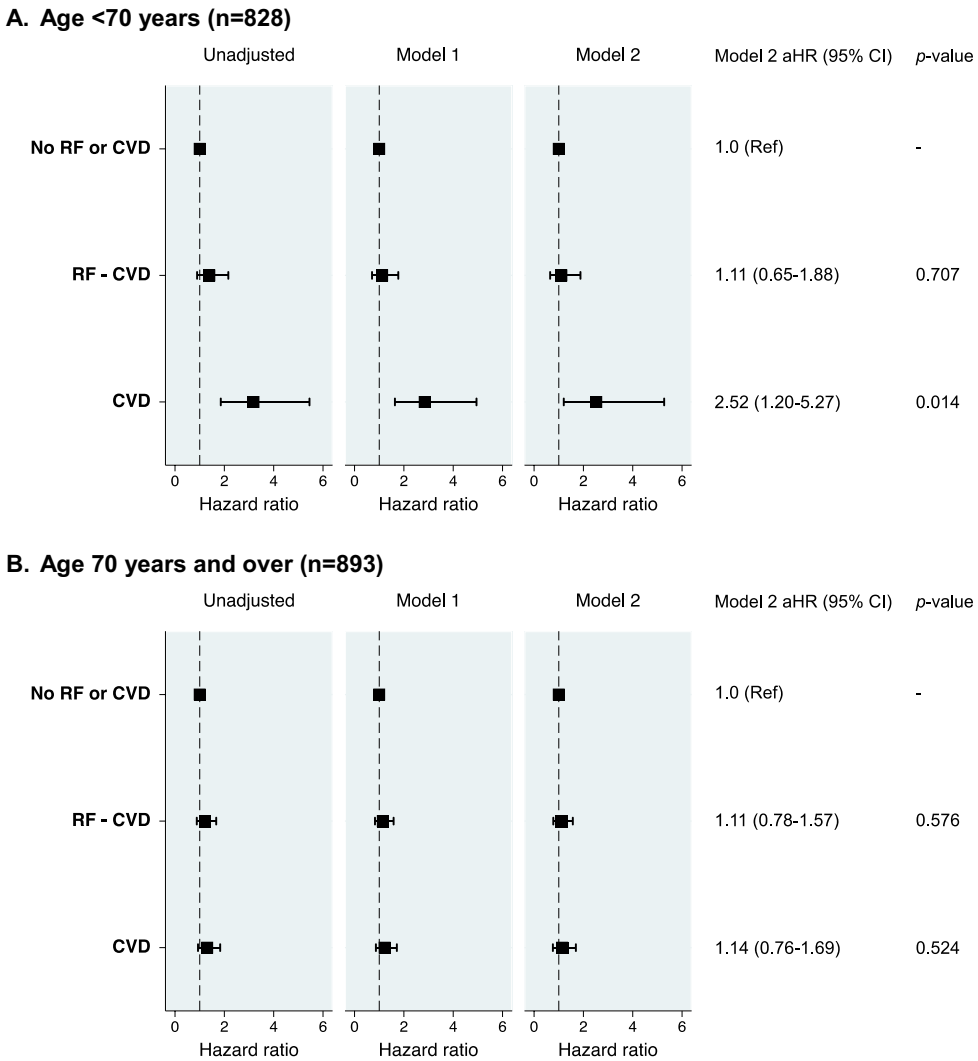


Fig. 2 Risk of in-hospital mortality stratified by age and cardiovascular risk group. **A** Age < 70 years (n = 828). **B**. Age 70 years and over (n = 893). aHR denotes adjusted hazard ratio; CVD, cardiovascular disease; RF, cardiovascular risk factors; RF-CVD, cardiovascular risk factors without established CVD. Model 1 adjusted for age, sex and ethnicity. Model 2 adjusted for age, sex, ethnicity, non-cardiac comorbidities (asthma, COPD, chronic renal failure, pulmonary embolism, DVT) and medications (ACEI or ARB, aldosterone receptor antagonists, beta-blockers, calcium channel blockers, loop diuretics, statins, anticoagulants, antiplatelet agents, metformin, sulphonylureas, SGLT2 inhibitors, DPP4 inhibitors, thiazolidinediones, GLP1 receptor agonists, insulin)

variable, (Additional file 3: Fig. S3A, B) or obesity as a categorical variable (Additional file 3: Fig. S4A, B) demonstrated similar effects to the main analysis. In addition, when obesity was included as a CV risk factor, 83 individuals with a BMI ≥ 30 kg/m^2 were reclassified as RF-CVD (from the no RF or CVD category). Effect estimates remained robust with only marginally wider confidence intervals (Additional file 3: Fig. S5A, B). Finally, a sensitivity analysis excluding current in-patients also showed similar findings to the main analysis. (Additional file 3: Fig. S6A, B).

Cardiovascular complications

Cardiovascular complications occurred in 431 (25.0%) patients, with two-thirds occurring in patients with CVD (n = 230, 65.9%). Patients with RF-CVD also had a higher CV complication rate than patients with neither CVD or CV risk factors (16.9% vs. 10.5%, Bonferroni adjusted $p < 0.001$). The most frequent CV complications were cardiac arrhythmias (84.7% atrial fibrillation), followed by acute heart failure (distinct from myocarditis) and acute MI, respectively (Table 2). Among patients presenting with an acute MI, 3 patients displayed ST elevation MI (STEMI) and underwent emergency percutaneous coronary intervention. Two additional patients underwent coronary angiography, one patient was diagnosed with spontaneous coronary artery dissection and one patient diagnosed with myocarditis. The remaining cases of acute MI were clinically considered to represent non-ST elevation or type 2 MIs [24]. The incidence of clinician-diagnosed myocarditis was low (0.7%). When arrhythmia-related complications were excluded, 59% of complications occurred in patients with CVD, 33% in patients with RF-CVD, and 8% in patients with neither.

In patients with CVD, the majority of CV complications represented exacerbations or decompensation of underlying CVD, rather than a new presentation, e.g., 86% of myocardial infarctions occurred in individuals with a previous myocardial infarction (Additional file 3: Fig. S7). Among specific CVDs and risk factors, pre-existing AF was associated with the highest adjusted odds of having any CV complication, followed by previous MI (Fig. 3A). For non-arrhythmia related CV complications, the highest adjusted odds were seen in patients with a previous myocardial infarction (Fig. 3B).

When CV complications were defined by cardiac biomarker elevation, in the subset of patients with a hs-cTnT measurement (n = 742), the presence of troponin elevation (Additional file 3: Fig. S8A) or troponin-elevation greater than 10 × normal (Additional file 3: Fig. S8B) were both associated with increased odds of in-hospital mortality across groups.

The incidence of VTE was higher in patients with CVD versus RF-CVD or neither (18.3% vs. 10.9% and 10.7% respectively, $p < 0.001$ for each). However, among CVDs and CV risk factors, hypertension (in the absence of established CVD) was the only independent CV predictor of VTE. Patients with previous VTE (n = 166, 48% anticoagulated at admission) had the highest rate of new (incident) VTE (49.4% vs. 8.4% with no previous VTE $p < 0.001$), and previous VTE was the strongest predictor of incident VTE, including adjustment for baseline anticoagulation use (Fig. 3C).

Discussion

We investigated the inter-relationship between CVD, CV risk factors, CV complications and mortality among 1721 consecutive patients hospitalised due to COVID-19. Overall, 20% of the cohort had CVD and an additional 50% had CV risk factors without yet having developed CVD (RF-CVD). A major finding is that the age- and sex-adjusted mortality risk is markedly increased in patients aged under 70 years with CVD but is only modestly and non-significantly increased in those with RF-CVD. The mortality risk associated with CVD appears much lower in individuals above 70 years of age. We also found that 1 in 4 patients hospitalised with COVID-19 experienced a CV complication, with cardiac arrhythmias representing the most common diagnosis, and the majority of CV complications and myocardial injury occurred in patients with CVD. Myocardial injury as indicated by an elevated troponin level was an independent predictor of mortality. Taken together, these findings suggest that pre-existing established CVD rather than CV risk factors per se influence mortality in severe COVID-19 and that this effect may be driven at least in part by CV complications and injury.

(See figure on next page.)

Fig. 3 Risk of COVID-19 related complications by cardiovascular risk group. **A** Any cardiovascular complications. **B** Non-arrhythmia related cardiovascular complications. **C** Venous thromboembolism. aOR, adjusted odds ratio; CV, cardiovascular; CVD, cardiovascular disease; DVT, deep vein thrombosis; PE, pulmonary embolism; MI, myocardial infarction; VTE, venous thromboembolism. Each model is adjusted for the variables listed, as well as age, sex, ethnicity, non-cardiac comorbidities (asthma, COPD, chronic renal failure, pulmonary embolism, DVT) and medications (ACEI or ARB, aldosterone receptor antagonists, beta-blockers, calcium channel blockers, loop diuretics, statins, anticoagulants, antiplatelet agents, metformin, sulphonylureas, SGLT2 inhibitors, DPP4 inhibitors, thiazolidinediones, GLP1 receptor agonists, insulin)

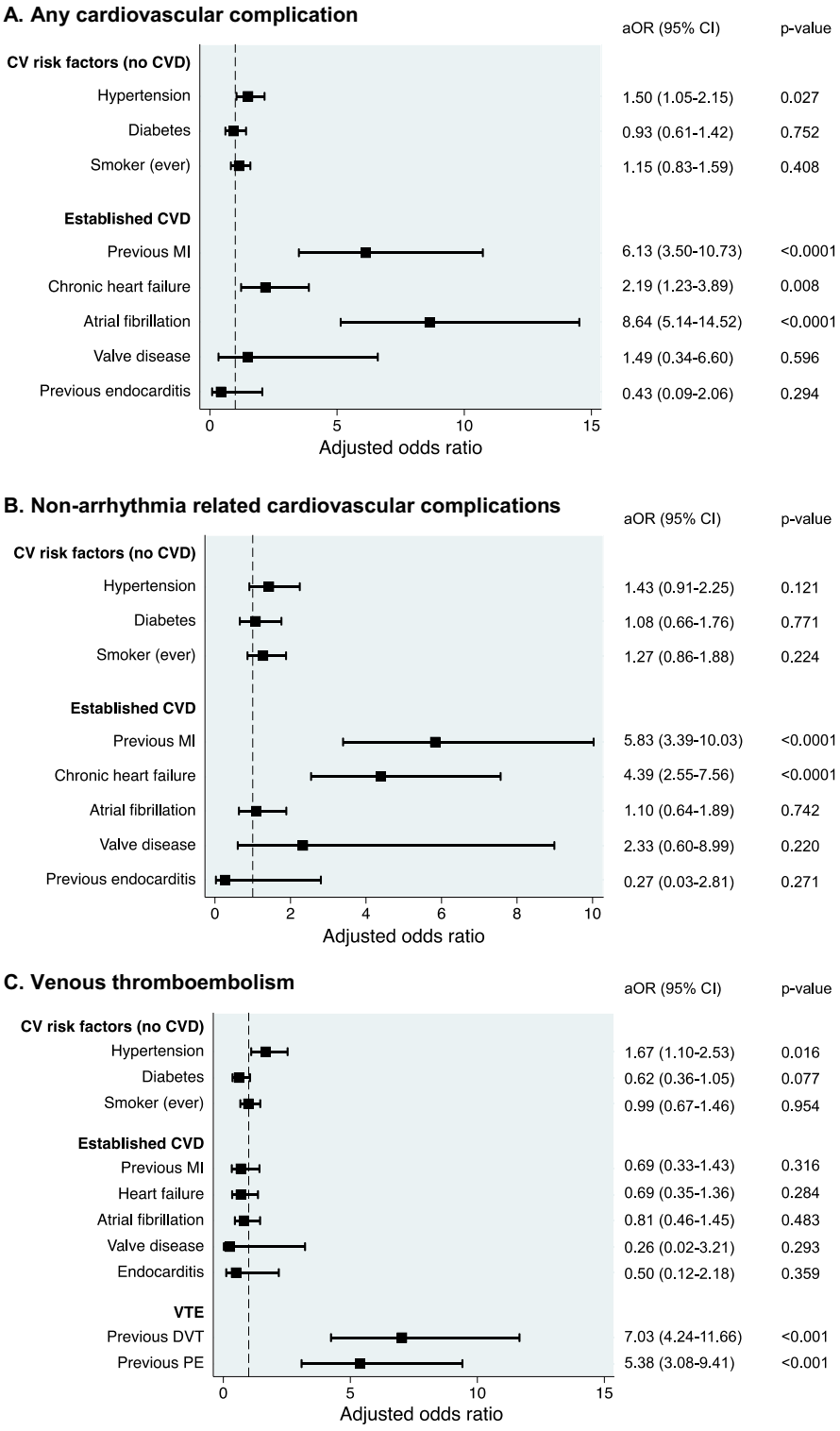


Fig. 3 (See legend on previous page.)

Pre-existing cardiovascular disease

The prevalence of CVD in our study was similar to other large hospital cohorts [4, 25]. We corroborate previous reports showing that a history of CVD is associated with greater risk of COVID-19-related mortality [4–6]. Interestingly, we identify an interaction with age, wherein the increased mortality risk is mainly apparent in people below 70 years of age whereas it is not statistically significant in people aged 70 and over. Reasons why older individuals do not also manifest higher CVD-related mortality warrant further investigation but may be related to a higher competing risk of non-CVD-related mortality due to frailty, non-CV comorbidities and immunosenescence, such that CVD has relatively minor additional prognostic effect. Similar age-dependent mortality effects have been reported in other studies [6, 7].

Cardiovascular risk factors

CV risk factors such as hypertension (56%) and diabetes (35%) were more prevalent than established CVD in our in-patient cohort, similar to other UK studies [4]. There are conflicting data regarding the prognostic impact of common CV risk factors in COVID-19. In hospitalised cohorts such as the UK ISARIC study, diabetes had a marginal independent effect on mortality risk, similar to our findings [4]. The majority of patients in ISARIC had uncomplicated diabetes and hypertension was not assessed. In the population-based OpenSAFELY study, diabetes was independently associated with a higher mortality, whereas hypertension was not independently associated with mortality [6]. Another large UK population-based study reported a 1.8-fold higher mortality risk for patients with type 2 diabetes after adjustment for relevant factors [26]. These divergent findings may be reconciled by considering that mortality rates in population-based studies reflect the risk of infection as well as risk of mortality once infected. For example, higher mortality risk associated with diabetes in population-based studies was suggested to be partly related to the level of glycaemic control [7]. We hypothesise that there may be an association between diabetes and infection risk. Previous studies have demonstrated that individuals with diabetes are at increased risk of serious infections [27] and poor glycaemic control has been associated with serious infections and hospital admission [28], although this association has not yet directly been shown for SARS-CoV-2. Nevertheless, even if patients with diabetes are more likely to be admitted, the current study and previous reports [4] suggest that mortality in this group could to a significant degree relate to the co-existence of CVD or other complications.

It should be acknowledged, however, that there may be interactions among CV risk factors and other variables that affect mortality risk e.g., ethnicity [7] and the effective treatment of risk factors [7].

Cardiovascular complications

A greater mortality risk associated with pre-existing CVD as compared to CV risk factors without CVD raises questions about the potential mechanisms underlying the higher risk. It has been proposed by several authors that endothelial dysfunction may be a major contributor to severe COVID-19 [29, 30]. Accordingly, pre-existing endothelial dysfunction may increase the likelihood of developing severe endothelial and vascular impairment with COVID-19. Such a mechanism would not readily explain the differential risk between established CVD and CV risk factors since both conditions are associated with endothelial dysfunction (and the majority of patients with CVD have CV risk factors). An alternative possibility is that patients with pre-existing established CVD are more prone to develop further cardiac injury and dysfunction which, in combination with pulmonary and right heart problems that represent the major manifestations of severe COVID-19, leads to life-threatening illness.

To explore this possibility, we analysed CV complications. The use of semi-automated pipelines to capture all clinician-diagnosed CV events minimised selection and indication bias. With this, our data demonstrate a high frequency of CV events overall (25%), with the majority (73%) representing arrhythmias, mostly atrial fibrillation, with rates comparable to smaller studies [5, 31, 32]. Other complications included acute MI and acute heart failure, with few clinically-diagnosed cases of myocarditis (0.7%).

Patients with pre-existing CVD had higher rates of CV complications than those with CV risk factors without CVD or patients without either CVD or CV risk factors. They also had higher rates of VTE but, whereas CVD was an independent predictor for incident CV complications, it was not a predictor of VTE. Importantly, we found that a high proportion of COVID-19 related CV complications (mainly cardiac) represent exacerbated or destabilised pre-existing CVD, rather than new presentations. Taken together, these findings suggest that the detrimental impact of pre-existing CVD on COVID-19 severity and mortality may be mediated mainly by increased cardiac problems rather than systemic vascular abnormalities such as VTE. In support of this idea, we found that myocardial injury as assessed by troponin elevation was most prevalent in patients with pre-existing CVD and strongly associated with mortality.

Currently the mechanism of thromboembolic risk in COVID-19 remains unclear. Potential mechanisms that have been suggested include vascular endothelial dysfunction, abnormal complement and coagulant pathway

activation, and abnormal platelet activation [33]. The fact that there was no independent association between new-onset VTE and established CVD may partly reflect higher rates of antiplatelet or anticoagulant therapy in this group. Another confounding factor could be undiagnosed VTE [10]. Ongoing clinical trials examining anticoagulation strategies and additional pathophysiology studies will provide further insights into this question.

The high incidence of cardiac arrhythmias, mostly atrial fibrillation, observed in this study may have multiple precipitants, such as myocardial ischaemia, increased sympathetic tone, inflammation (systemic as well as myocardial), and electrolyte imbalance. Our finding of a low incidence of myocarditis is consistent with several other reports and a recent review of autopsy cases [34, 35].

Clinical implications

Our finding that a large proportion of CV complications represented destabilised pre-existing CVD, supports the importance of identifying CVD in patients presenting to hospital with COVID-19 (including new diagnoses) and maintaining evidence-based CV care, alongside disease-specific treatment for COVID-19, including, for example, continuation of ACE inhibitors or ARBs in individuals with an indication [16]. The high incidence of arrhythmias may warrant more systematic electrocardiographic screening of hospitalised patients, particularly as the detection of new-onset atrial fibrillation is an indication for anticoagulation to reduce the risk of stroke and systemic thromboembolism. The low rates of STEMI in our cohort, matches reports of declining admission rates for STEMI during the pandemic [36], suggesting that clinical deterioration and CV complications in patients with COVID-19 may be less frequently due to STEMI. Additionally, given the age and comorbidity profile of patients hospitalised with severe COVID-19, this population has a high risk of type 2 myocardial infarction (i.e., a mismatch between oxygen supply and demand, without acute atherothrombotic plaque disruption [24]), in the presence or absence of underlying CAD. Since this may have implications for triage and treatment protocols, a low threshold for biomarker (troponin) assessment in patients with pre-existing CVD should be considered.

Limitations

Our analysis was limited to individuals who required hospital admission and is therefore only generalisable to this population. This was a retrospective study of prospectively entered data in the EHR. Although this study assessed CV risk factors and established CVD as separate entities, a limitation of this approach is that there is a continuum and that individuals with risk factors may have undiagnosed CVD. Nevertheless, the presence of overt diagnosed CVD does appear to distinguish this group in terms of outcomes. During the early stages of the pandemic, echocardiography and coronary angiography were only performed in selected cases in keeping with recommendations to avoid unnecessary cardiac imaging (in order to reduce transmission of the virus, protect healthcare professionals, and conserve personal protective equipment [37]). A small minority of patients were still in hospital and were censored at the study end-date (2.1%). However, a sensitivity analysis in patients who were either discharged or died revealed similar findings (Additional file 3: Fig. S6). Our selection of cardiovascular risk factors was based on those with highest prevalence, most reliably reported, and to further explore findings in preceding large UK population studies [4, 7]. We did not have robust information regarding dyslipidaemia, however a sensitivity analysis reclassifying patients prescribed statin therapy (with no known CVD) as RF-CFD, as a crude measure of hyperlipidaemia or high CV risk, showed similar effect estimates to the main analysis. In addition, there was a significant amount of missing data for BMI, however effect estimates were robust in sensitivity analyses accounting for BMI and obesity, including obesity as a RF-CVD. We considered this approach more appropriate than multiple imputation, because underweight and overweight individuals, may be more likely to have their BMI recorded, thus contradicting the required missing at random assumption [38].

Finally, our multivariable analyses adjusted for patient characteristics and the presence or absence of several comorbidities, however measures of control (e.g., blood pressure control for hypertension or HbA1c for diabetes) were not assessed and may impact the risk of death. Additionally, we cannot exclude residual confounding due to unmeasured non-cardiac comorbidities, such as malignancy, hepatic or non-vascular neurological diseases.

Conclusions

Among patients hospitalised with COVID-19, pre-existing established CVD appears to be a more important contributor to in-hospital mortality than CV risk factors without co-existent CVD, particularly in patients below the age of 70 years. This enhanced risk may be driven, at least in part, by a higher incidence of cardiac complications and myocardial injury in patients with pre-existing CVD whereas VTE appears less important. Optimal management of pre-existing CVD may serve to modify outcomes related to COVID-19 in this group. In addition, heightened vigilance for arrhythmias and myocardial injury should be considered for patients with pre-existing CVD, to enable early detection and intervention where needed.

Abbreviations
ACEI: Angiotensin converting enzyme inhibitors; aHR: Adjusted hazard ratio; ARB: Angiotensin receptor blockers; ARDS: Acute respiratory distress syndrome; BMI: Body mass index; CKD: Chronic kidney disease; COPD: Chronic obstructive pulmonary disease; COVID-19: Coronavirus disease 2019; CV: Cardiovascular; CVD: Cardiovascular disease; DPP4: Dipeptidyl-peptidase 4; DVT: Deep vein thrombosis; ESC: European Society of Cardiology; EHR: Electronic health record; GLP1: Glucagon-like peptide 1; HR: Hazard ratio; hs-cTnT: High sensitivity cardiac troponin T; ICU: Intensive care unit; IQR: Interquartlie range (25th–75th percentile); NHS: National Health Service; MI: Myocardial infarction; NLP: Natural language processing; PE: Pulmonary embolism; RF: Cardiovascular risk factors; RF-CVD: Cardiovascular risk factors without established cardiovascular disease; RT-PCR: Reverse transcriptase polymerase chain reaction; SARS-CoV-2: Severe acute respiratory distress syndrome-coronavirus-2; SGLT2: Sodium glucose co-transporter-2; STEMI: S-T elevation myocardial infarction; SD: Standard deviation; VTE: Venous thromboembolism.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12872-021-02137-9.

Additional file 1. Supplemental Methods.

Additional file 2. Supplemental Table 1.

Additional file 3. Supplemental Figures.

Acknowledgements
We are extremely grateful to all the clinicians and NHS staff who have looked after our patients with COVID-19. We additionally thank the patient experts of the KERRI committee.

Authors' contributions
RZ, KOG, and AMS conceived the study. KOG, AS, AMS and RZ participated in the study design. KOG, AS, DB, and RZ participated in the data collection. RZ performed the data analyses. KOG, RB, AMS and RZ contributed to data interpretation. KOG, AMS and RZ drafted the first version of the manuscript. All authors contributed to and approved the final manuscript and the decision to submit. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. All authors read and approved the final manuscript.

Funding
This work was supported in part by the British Heart Foundation (CH/1999001/11735 and RE/18/2/34213 to AMS) and the National Institute for Health Research Biomedical Research Centres (NIHR BRCs) at Guy's & St Thomas' NHS Foundation Trust (IS-BRC-1215-20006) and South London and Maudsley NHS Foundation Trust (SLAM; IS-BRC-1215-20018) both with King's College London. RZ is also supported by a King's Prize Fellowship. AMS is also supported by the Fondation Leducq. KO'G is supported by a MRC Clinical Training Fellowship. AS is funded by the King's Medical Research Trust. DMB holds a UKRI Fellowship as part of HDRUK MR/S00310X/1. RB is supported by a Medical Research Council (MRC) Skills Development Fellowship programme (MR/R016372/1) and the NIHR SLAM BRC. RJBD is also supported by Health Data Research UK (HDRUK); UK Research and Innovation (UKRI) London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; the BigData@Heart Consortium (Grant No. 116074 of the European Union Horizon 2020 programme); the NIHR BRC and Research Informatics Unit at University College London Hospitals; and the NIHR Applied Research Collaboration South London at KCHFT. The views expressed are those of the authors and not necessarily those of NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The authors declare that all data supporting the findings of this study are available within the article (and its supplementary information files). Individual participant data will not be made available due to confidentiality regulations.

Declarations

Ethics approval and consent to participate
This study was conducted under London South East Research Ethics Committee approval (reference 18/LO/2048) granted to the King's Electronic Records Research Interface (KERRI) for collection of clinically relevant data from patients' electronic health records. Individual patient written consent was not required, however, COVID-19 work was reviewed with expert patient input on a virtual committee with Caldicott Guardian oversight.

Consent for publication
Not applicable.

Competing interests
JTHT received research funding from Innovate UK & Office of Life Sciences, and iRhythm Technologies, and holds shares < £5,000 in Glaxo Smithkline and Biogen. The other authors declare no competing interests.

Author details
[1]Department of Cardiology, King's College London British Heart Foundation Centre of Research Excellence, School of Cardiovascular Medicine and Sciences, London, UK. [2]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [3]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK. [4]King's College Hospital NHS Foundation Trust, London, UK. [5]Health Data Research UK London, Institute of Health Informatics, University College London, London, UK. [6]School of Cardiovascular Medicine and Sciences, James Black Centre, King's College London, 125 Coldharbour Lane, London SE5 9NU, UK.

Received: 7 March 2021 Accepted: 24 June 2021
Published online: 03 July 2021

References

1. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis. 2020;94:91–5.
2. Guan WJ, Liang WH, Zhao Y, Liang HR, Chen ZS, Li YM, et al. Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. Eur Respir J. 2020;55(5).
3. Zakeri R, Bendayan R, Ashworth M, Bean DM, Dodhia H, Durbaba S, et al. A case-control and cohort study to determine the relationship between ethnic background and severe COVID-19. EClinicalMedicine. 2020:100574.
4. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020;369:m1985.
5. Inciardi RM, Adamo M, Lupi L, Cani DS, Di Pasquale M, Tomasoni D, et al. Characteristics and outcomes of patients hospitalized for COVID-19 and cardiac disease in Northern Italy. Eur Heart J. 2020;41(19):1821–9.
6. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430–6.
7. Holman N, Knighton P, Kar P, O'Keefe J, Curley M, Weaver A, et al. Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: a population-based cohort study. Lancet Diabetes Endocrinol. 2020;8(10):823–33.
8. NHS. https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/whos-at-higher-risk-from-coronavirus/ 2020 [
9. Schaller T, Hirschbuhl K, Burkhardt K, Braun G, Trepel M, Markl B, et al. Postmortem examination of patients with COVID-19. JAMA. 2020;323(24):2518–20.
10. Wichmann D, Sperhake JP, Lutgehetmann M, Steurer S, Edler C, Heinemann A, et al. Autopsy Findings And Venous Thromboembolism In Patients With COVID-19: a prospective cohort study. Ann Intern Med. 2020;173(4):268–77.

11. Hu H, Ma F, Wei X, Fang Y. Coronavirus fulminant myocarditis saved with glucocorticoid and human immunoglobulin. Eur Heart J. 2020.
12. Hua A, O'Gallagher K, Sado D, Byrne J. Life-threatening cardiac tamponade complicating myo-pericarditis in COVID-19. Eur Heart J. 2020;41(22):2130.
13. Sanchez-Recalde A, Solano-Lopez J, Miguelena-Hycka J, Martin-Pinacho JJ, Sanmartin M, Zamorano JL. COVID-19 and cardiogenic shock. Different cardiovascular presentations with high mortality. Rev Esp Cardiol (Engl Ed). 2020;73(8):669–72.
14. Guo T, Fan Y, Chen M, Wu X, Zhang L, He T, et al. Cardiovascular Implications of Fatal Outcomes of Patients With Coronavirus Disease 2019 (COVID-19). JAMA Cardiol. 2020.
15. Lala A, Johnson KW, Januzzi JL, Russak AJ, Paranjpe I, Richter F, et al. Prevalence and Impact of Myocardial Injury in Patients Hospitalized With COVID-19 Infection. J Am Coll Cardiol. 2020;76(5):533–46.
16. Bean DM, Kraljevic Z, Searle T, Bendayan R, Kevin O, Pickles A, et al. Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. Eur J Heart Fail. 2020;22(6):967–74.
17. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack—experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Med Inform Decis Mak. 2018;18(1):47.
18. Bean DM, Teo J, Wu H, Oliveira R, Patel R, Bendayan R, et al. Semantic computational analysis of anticoagulation use in atrial fibrillation from real world data. PLoS ONE. 2019;14(11):e0225625.
19. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean DM, et al. Multi-domain Clinical Natural Language Processing with MedCAT: the Medical Concept Annotation Toolkit. Preprint: https://arxivorg/abs/201001165. 2020.
20. Searle T, Z K, Bendayan R, Bean D, R. D. In: Proceedings of the 2019 conference on natural language processing. 2019:139–44.
21. Baumgartner H, Falk V, Bax JJ, De Bonis M, Hamm C, Holm PJ, et al. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. Eur Heart J. 2017;38(36):2739–91.
22. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18(6):681–94.
23. Grambsch PM. Goodness-of-fit and diagnostics for proportional hazards regression models. Cancer Treat Res. 1995;75:95–112.
24. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, et al. Fourth universal definition of myocardial infarction (2018). Circulation. 2018;138(20):e618–51.
25. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. JAMA. 2020;323(20):2052–9.
26. Barron E, Bakhai C, Kar P, Weaver A, Bradley D, Ismail H, et al. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. Lancet Diabetes Endocrinol. 2020;8(10):813–22.
27. Carey IM, Critchley JA, DeWilde S, Harris T, Hosking FJ, Cook DG. Risk of infection in type 1 and type 2 diabetes compared with the general population: a matched cohort study. Diabetes Care. 2018;41(3):513–21.
28. Critchley JA, Carey IM, Harris T, DeWilde S, Hosking FJ, Cook DG. Glycemic control and risk of infections among people with type 1 or type 2 diabetes in a large primary care cohort study. Diabetes Care. 2018;41(10):2127–35.
29. Libby P, Luscher T. COVID-19 is, in the end, an endothelial disease. Eur Heart J. 2020;41(32):3038–44.
30. Gustafson D, Raju S, Wu R, Ching C, Veitch S, Rathnakumar K, et al. Overcoming barriers: the endothelium as a linchpin of coronavirus disease 2019 pathogenesis? Arterioscler Thromb Vasc Biol. 2020;40(8):1818–29.
31. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA. 2020.
32. Peltzer B, Manocha KK, Ying X, Kirzner J, Ip JE, Thomas G, et al. Outcomes and Mortality Associated with Atrial Arrhythmias Among Patients Hospitalized with COVID-19. J Cardiovasc Electrophysiol. 2020.
33. Loo J, Spittle DA, Newnham M. COVID-19, immunothrombosis and venous thromboembolism: biological mechanisms. Thorax. 2021.
34. Halushka MK, Vander Heide RS. Myocarditis is rare in COVID-19 autopsies: cardiovascular findings across 277 postmortem examinations. Cardiovasc Pathol. 2020;50:107300.
35. Kawakami R, Sakamoto A, Kawai K, Gianatti A, Pellegrini D, Nasr A, et al. Pathological evidence for SARS-CoV-2 as a cause of myocarditis: JACC review topic of the week. J Am Coll Cardiol. 2021;77(3):314–25.
36. Rattka M, Dreyhaupt J, Winsauer C, Stuhler L, Baumhardt M, Thiessen K, et al. Effect of the COVID-19 pandemic on mortality of patients with STEMI: a systematic review and meta-analysis. Heart. 2020.
37. Cosyns B, Lochy S, Luchian ML, Gimelli A, Pontone G, Allard SD, et al. The role of cardiovascular imaging for myocardial injury in hospitalized COVID-19 patients. Eur Heart J Cardiovasc Imaging. 2020;21(7):709–14.
38. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, 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.
39. Kellum J, Lameire, N, Aspelin, P, Barsoum, RS, Burdmann, EA, Goldstein, SL, Herzog, CA, Joannidis, M, Kribben, A, Levey, AS, MacLeod, AM, Mehta, RL, Murray, PT, Naicker, S, Opal, SM, Schaefer, F, Schetz, M & Uchino, S. Kidney disease: Improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl. 2012;2(1):1–138.
40. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J. 2016;37(27):2129–200.
41. Caforio AL, Pankuweit S, Arbustini E, Basso C, Gimeno-Blanes J, Felix SB, et al. Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. Eur Heart J. 2013;34(33):2636–48, 48a–48d.
42. Adler Y, Charron P, Imazio M, Badano L, Baron-Esquivias G, Bogaert J, et al. 2015 ESC Guidelines for the diagnosis and management of pericardial diseases: the task force for the diagnosis and management of pericardial diseases of the European Society of Cardiology (ESC) endorsed by: the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. 2015;36(42):2921–64.
43. Habib G, Lancellotti P, Antunes MJ, Bongiorni MG, Casalta JP, Del Zotti F, et al. 2015 ESC Guidelines for the management of infective endocarditis: the task force for the management of infective endocarditis of the European Society of Cardiology (ESC). Endorsed by: European Association for Cardio-Thoracic Surgery (EACTS), the European Association of Nuclear Medicine (EANM). Eur Heart J. 2015;36(44):3075–128.

Publisher's Note

7.2 Machine Learning Enabled Multi-Trust Audit of Stroke Comorbidities Using Natural Language Processing

7.2.1 Authors

Shek, Anthony; Jiang, Zhilin; Teo, James; Au Yeung, Joshua; Bhalla, Ajay; Richardson, M.P.; Mah, Yee

This work was published in the journal: European journal of neurology
DOI: 10.1111/ene.15071

Statement of Contributions

Conceptualization: YM;

Data curation: AS, ZJ, YM, JT, JY, AB;

Formal analysis YM, AS, JT;

Supervision: YM, JT, MR;

Writing-review: All

Received: 7 August 2021 | Accepted: 10 August 2021

DOI: 10.1111/ene.15071

european journal of neurology
the official journal of the european academy of neurology

ORIGINAL ARTICLE

Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing

Anthony Shek[1] | Zhilin Jiang[2] | James Teo[1,2] | Joshua Au Yeung[3] | Ajay Bhalla[3] | Mark P. Richardson[1,2,4] | Yee Mah[1,5]

[1]School of Neuroscience, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, London, UK

[2]King's College Hospital NHS Foundation Trust, London, UK

[3]Guy's & St Thomas' Hospital NHS Foundation Trust, London, UK

[4]National Institute for Health Research, Maudsley Biomedical Research Centre, South London & Maudsley NHS Foundation Trust, London, UK

[5]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

Correspondence
Yee Mah, School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK.
Email: yee.mah@nhs.net

Funding information
A.S. is supported by the Kings Medical Research Trust. Y.M. is supported by an MRC Clinical Academic Research Partnership grant (MR/T005351/1).

Abstract

Background and purpose: With the increasing adoption of electronic records in the health system, machine learning-enabled techniques offer the opportunity for greater computer-assisted curation of these data for audit and research purposes. In this project, we evaluate the consistency of traditional curation methods used in routine clinical practice against a new machine learning-enabled tool, MedCAT, for the extraction of the stroke comorbidities recorded within the UK's Sentinel Stroke National Audit Programme (SSNAP) initiative.

Methods: A total of 2327 stroke admission episodes from three different National Health Service (NHS) hospitals, between January 2019 and April 2020, were included in this evaluation. In addition, current clinical curation methods (SSNAP) and the machine learning-enabled method (MedCAT) were compared against a subsample of 200 admission episodes manually reviewed by our study team. Performance metrics of sensitivity, specificity, precision, negative predictive value, and F1 scores are reported.

Results: The reporting of stroke comorbidities with current clinical curation methods is good for atrial fibrillation, hypertension, and diabetes mellitus, but poor for congestive cardiac failure. The machine learning-enabled method, MedCAT, achieved better performances across all four assessed comorbidities compared with current clinical methods, predominantly driven by higher sensitivity and F1 scores.

Conclusions: We have shown machine learning-enabled data collection can support existing clinical and service initiatives, with the potential to improve the quality and speed of data extraction from existing clinical repositories. The scalability and flexibility of these new machine-learning tools, therefore, present an opportunity to revolutionize audit and research methods.

KEYWORDS
clinical coding, natural language processing, programme evaluation, stroke

INTRODUCTION

Medical records are a rich source of information, continuously accessed by health care professionals to help care for their patients and community. The benefits of trawling through swathes of medical notes are clear, including understanding the individual in the acute setting; audit and service evaluation [1–3]; and identifying patterns embedded in a disease population for research [4–6]. With the increasing adoption of electronic records in the health system [7–10], using computers to analyse all these data has been a common objective [11–13]. However, accurate extraction of medical concepts from unstructured data, like free text, requires an understanding of the language used, something that is relatively simple for a human but extremely challenging for a computer.

Over the past decade advancements in a branch of machine learning, known as natural language processing (NLP), have enabled the translation of free text into a standardized, structured set of medical terms that can be subsequently analysed by a computer [14]. These tools have the potential to automate and support data collection; however, evaluation with real-world clinical data across medical specialties, such as stroke, has been limited [15]. CogStack is an open-source software ecosystem that incorporates both the structured and unstructured components of the electronic health record (EHR). The MedCAT Toolkit [16] component supports the development of NLP algorithms through the ability to disambiguate and capture synonyms, and acronyms for medical Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) concepts, as well as surrounding linguistic context such as negation, subject, and basic grammatical tense, using deep learning and long short-term memory networks. Further supervised training can improve the detection of annotations and meta-annotations using the MedCAT Trainer platform. The entire CogStack ecosystem is open source and available on GitHub (https://github.com/CogStack).

Conventional registry and national audits use standardized case report forms to provide periodic standardized submissions into centralized databases. The Sentinel Stroke National Audit Programme (SSNAP) [17] is a health care quality improvement programme collecting stroke patient data that represent >90% of all cases in England, Wales, and Northern Ireland. With 100,000 stroke cases per annum [17], this is a time-pressured, labour-intensive exercise conducted manually by a team of clinical coders and/or clinicians. Although manual curation is the current gold standard, these pressures increase the risk of errors [18–20] and limit the timeliness of the data to some months after the event, negatively impacting on the utility of the collected data.

In this project, we evaluated the manually inputted SSNAP data from three different National Health Service (NHS) hospitals against a manually reviewed sample for the four stroke comorbidities routinely collected as part of the SSNAP initiative. We also trained MedCAT on a set of manually annotated stroke documents, to identify the same four comorbidities, and then applied the model to the inpatient stroke records of three different NHS hospitals, comparing the MedCAT performances against the corresponding manually inputted SSNAP data and the manually reviewed subsample.

METHODS

Datasets

All admission episodes contained within SSNAP were identified for the following three NHS hospitals in the time period 1 January 2019 to 1 April 2020: King's College Hospital NHS Foundation Trust (KCH); Princess Royal University Hospital NHS Trust (PRUH); and Guy's and St Thomas' NHS Foundation Trust (GSTT). KCH and PRUH share a common EHR system, whereas GSTT uses a clinically separate electronic documentation system. A total of 2327 admission episodes were included in this study. Table 1 shows the breakdown of episodes. Patients without any available electronic notes were excluded from the study.

This project was conducted under audit and data processing for service evaluation. Research ethics review was not required. For

TABLE 1 Table showing the distribution of patient episodes and comorbidities for the different datasets

	KCH-PRUH	KCH subsample	*p*	GSTT	GSTT subsample	*p*
Number of episodes	2136	100		191	100	
Excluded episodes	124	–		18	–	
Female	1066 (49.9%)	51		96 (50.3%)	47	
Mean age, years	71.5 (SD = 15.1)	66.9 (SD = 16.0)	0.58	70.9 (SD = 14.4)	71.3 (SD = 12.5)	0.94
AIS	1864 (87.3%)	88 (88%)	0.52	166 (86.9%)	89 (89%)	0.08
AF	446 (20.9%)	21 (21%)	0.88	35 (18.3%)	15 (15%)	0.79
Hypertension	1235 (57.8%)	60 (60%)	0.97	109 (57.1%)	62 (62%)	0.91
CCF	90 (4.2%)	8 (8%)	0.08	12 (6.3%)	7 (7%)	0.40
Diabetes	520 (24.3%)	32 (32%)	0.63	59 (30.1%)	32 (32%)	0.44

Note: Patient episodes without digital documents were excluded. A Wilcoxon rank-sum test was performed to assess for differences between the subsample and parent sample for KCH-PRUH and GSTT.

Abbreviations: AF, atrial fibrillation; AIS, acute ischaemic stroke; CCF, congestive cardiac failure; GSTT, Guy's and St Thomas' Hospital; KCH, King's College Hospital; PRUH, Princess Royal University Hospital.

transparency and conformance with Caldicott Principle 8 [21], a patient–public research oversight group - King's Electronic Records Research Initiative - has provided positive feedback on projects that use natural language processing for clinical data processing for coding and service evaluation. The data that support the findings of this study are available on request and review from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

MedCAT algorithm training and data

The base MedCAT algorithm was trained in an unsupervised manner on the entire KCH EHR data consisting of more than 18 million documents [16], and received additional training from 301 and 373 annotated documents in endocrinology and cardiology, respectively. For our study, further training on stroke-specific comorbidities was provided through 500 KCH-PRUH annotated documents obtained from 2015 to October 2020, stratified by patient, age, and gender, using the method described in Kraljevic et al. [16]. This only included free-text information documented by clinical staff, and excluded information from other systems like blood results, investigation reports, outpatient letters, and vital observations. MedCAT counted the number of instances a concept was mentioned (e.g., atrial fibrillation [AF]) and generated a total count for each patient episode. This only included references for the presence of the concept relating to a patient. Phrases such as “this patient does not have AF” or “a family history of AF” would not increase the count. Because MedCAT is mapped onto the SNOMED-CT library, counts for child concepts defined by the inbuilt “IS A” hierarchical relationship were merged to reflect the data collected in SSNAP. Table S3 shows the list of SNOMED-CT concepts used to emulate the SSNAP comorbidities.

The MedCAT concept count was converted to a binary state by applying a threshold, above which a patient would be diagnosed with the comorbidity for the specific admission episode. Two different document periods were examined, based on the recorded admission and discharge timestamps: (i) 12 h prior to admission to 12 h after discharge (admission period) and (ii) January 2015 to 12 h after discharge (2015-to-discharge).

SSNAP data

The SSNAP governing body has released protocols and guidelines for data curation, with each participating site responsible for its own curation of data [17]. Although the data collected for SSNAP have evolved with the changing face of stroke, the data definition for SSNAP remained constant during the period assessed in our study.

Using the local SSNAP data from each hospital, the comorbidities AF, hypertension, congestive cardiac failure (CCF), and diabetes were extracted. SSNAP collects both “atrial fibrillation” and “new atrial fibrillation,” where the patient cannot be positive for both labels. For this project, we combined these two groups to represent whether the comorbidity of AF was present for this admission episode and used this to perform the subsequent analyses. Diabetes included both Type 1 and Type 2 diabetes mellitus. In addition, the stroke type for the admission episode—acute ischaemic stroke (AIS) or primary intracerebral haemorrhage (PIH)—was recorded.

Subsample reread (KCH and GSTT): Ground truth

To evaluate the performance of the two auditing methods, a mutual reference dataset was curated to represent the ground truth. Two subsamples of 100 patient episodes each were randomly selected from the KCH and GSTT datasets. A Wilcoxon rank-sum test was used to assess whether there were significant differences between the subsample reread and its parent dataset.

To curate the subsample reread, the range of documents and level of access were identical to what would have been available to the SSNAP operators. In contrast to the SSNAP curation method, the subsample reread was collected solely by trained clinicians (Y.M., J.T., A.B., Z.J., J.A.Y.) who were not under time constraints to extract the comorbidities. The medical notes for each admission episode were reviewed, excluding any documents that were created after the episode discharge date. The presence or absence of AF, hypertension, CCF, and diabetes was noted. They also recorded whether the acute admission was either an AIS or a PIH. For each site, the level of interrater agreement for the four comorbidities was assessed using Cohen kappa on 20% of the sample. In cases of disagreement between operators, the finding from the more senior clinician was used.

Evaluation

For purposes of brevity, SSNAP is used to indicate data held in SSNAP that are obtained through manual data curation. MedCAT is used to indicate data generated by NLP from hospital source documents. A series of comparisons were then performed to evaluate the two auditing methods using the three different datasets: MedCAT, SSNAP, and subsample reread (Figure 1).

Two sets of comparisons were conducted: first, MedCAT was assessed against SSNAP data for the two different EHR systems (Figure 1, comparison A); second, both MedCAT and SSNAP were assessed against the ground truth, subsample reread of KCH ($n = 100$) and GSTT ($n = 100$) datasets (Figure 1, comparisons B and C). Metrics of sensitivity, specificity, precision, negative predictive value, and F1 score were calculated. The level of agreement between the subsample reread and the auditing methods was measured using Cohen kappa, whereas the McNemar test was used to assess the difference in performance between the auditing methods. All statistical analyses were performed using MATLAB 2020b [22].

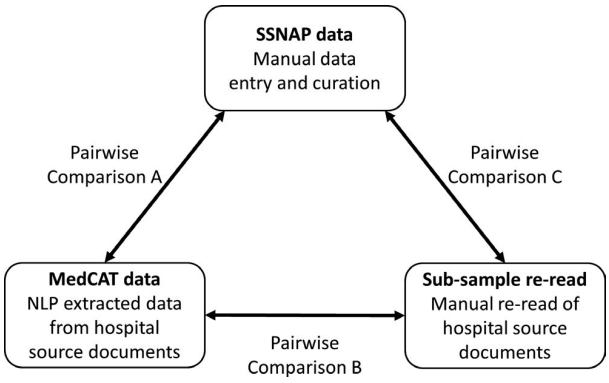


FIGURE 1 Diagram illustrating the three different pairwise comparisons used in the study. Sentinel Stroke National Audit Programme (SSNAP) data were manually curated; MedCAT refers to the machine learning-enabled extraction of data; and the subsample reread (ground truth) refers to the patient episodes manually reviewed again for the study. NLP, natural language processing

RESULTS

The mean ages for the KCH-PRUH and GSTT datasets were 71.5 and 70.9 years, respectively. There was no significant difference between the subsample reread and the respective parent dataset for age, proportion of females, AIS, and comorbidities. The absolute values for each dataset are displayed in Table 1. There was a significantly higher proportion of females at the GSTT site compared with KCH-PRUH ($p = 0.045$). The prevalence of comorbidities between the KCH-PRUH and GSTT sites was not significantly different except for diabetes mellitus ($p = 0.045$).

The F1 score is the harmonic mean between the sensitivity and precision (positive predictive value). For the KCH-PRUH and GSTT datasets, compared against SSNAP, MedCAT was able to determine the type of stroke using documents from 2015-to-discharge with peak F1 scores of 0.92 and 0.95, respectively (Table S2).

Comparison A: MedCAT compared against SSNAP

Comparing MedCAT against SSNAP, the performance of MedCAT is similar between the two document inclusion periods, with the peak F1 scores obtained within a threshold value between two and eight counts (Table S1). MedCAT's performance as a function of MedCAT count threshold for the type of stroke, and the four comorbidities are provided in Figures S1 and S2. The corresponding area under the receiver operating characteristic curve plots for the four comorbidities are displayed in Figure S3.

The peak F1 scores for AF and diabetes were obtained using documents from the admission period only, whereas CCF was from all documents (2015-to-discharge), and there was no difference with hypertension (Table S1). The deterioration in F1 score for AF in the GSTT data compared with the KCH-PRUH data is primarily driven by a low level of precision (i.e., false positives) likely related to the number of acronyms for atrial fibrillation (e.g., "AF," "PAF," "AFib," "atrial fib") not encountered in the training sample.

Comparison B and C: MedCAT/SSNAP compared against the subsample reread (ground truth)

Interrater agreement for the two sites were high, with a Cohen kappa of 0.89 for diabetes, and 1.0 for the other three comorbidities at the KCH site, and 0.83 for hypertension and CCF, and 1.0 for AF and diabetes at the GSTT site.

To facilitate the comparison, a threshold heuristic of five counts was selected for the MedCAT models used at each site and applied to document inclusion periods of admission only and

TABLE 2 Table showing the performance metrics of SSNAP and MedCAT against the manually reviewed (ground truth) subset

	Sensitivity		Specificity		Precision		NPV		F1 score	
	SSNAP	MedCAT	SSNAP	MedCAT	SSNAP	MedCAT	SSNAP	MedCAT	SSNAP	MedCAT
KCH										
AF	0.72	1.0	1.0	0.92	1.0	0.83	0.90	1.0	0.84	0.91
HTN	0.84	0.91	0.97	0.77	0.98	0.9	0.72	0.79	0.91	0.91
CCF	0.39	0.56	0.99	1.0	0.88	1.0	0.88	0.91	0.54	0.71
DM	0.84	0.97	1.0	0.94	1.0	0.9	0.91	0.98	0.91	0.94
GSTT										
AF	0.5	0.89	1.0	0.87	1.0	0.74	0.84	0.94	0.67	0.81
HTN	0.72	0.91	0.92	0.89	0.96	0.96	0.50	0.73	0.82	0.93
CCF	0.36	0.50	0.97	0.99	0.63	0.88	0.94	0.95	0.45	0.64
DM	0.76	0.91	0.98	0.94	0.95	0.89	0.87	0.97	0.84	0.90

Note: The MedCAT model used a threshold heuristic of 5 and was applied to all available documents until discharge (2015-to-discharge).

Abbreviations: AF, atrial fibrillation; CCF, congestive cardiac failure; DM, diabetes mellitus; GSTT, Guy's and St Thomas' Hospital; HTN, hypertension; KCH, King's College Hospital; NPV, negative predictive value; SSNAP, Sentinel Stroke National Audit Programme.

2015-to-discharge. In the KCH dataset, SSNAP obtained F1 scores of 0.84, 0.91, and 0.91 for AF, hypertension, and diabetes, with CCF performing worse, with a score of 0.54. In comparison, MedCAT achieved F1 scores greater than or equal to 0.91 in all but CCF, with a score of 0.71 (Table 2). The F1 scores for MedCAT were higher when using all available documents until discharge (2015-to-discharge; Table S2). Similarly, in the GSTT dataset, SSNAP obtained F1 scores of 0.67, 0.81, and 0.85 for AF, hypertension, and diabetes, with a lower score of 0.40 for CCF (Table 2). MedCAT obtained F1 scores greater than 0.8 for all comorbidities apart from CCF, which was 0.64. Unlike the KCH dataset, the F1 scores were not consistently higher when using all available documents, compared with only those from the admission period, with the reverse found for AF and diabetes (Table S2).

The MedCAT audit method achieved substantial levels of agreement with the subsample reread at both sites (Table 3), with near-perfect agreement achieved with AF at KCH, and diabetes at both sites. In contrast, SSNAP achieved a lower level of agreement for all comorbidities at both sites, except for hypertension at KCH. The MedCAT audit method was significantly superior to SSNAP for hypertension in the GSTT dataset, and almost reached significance for CCF in both the KCH and GSTT datasets, with a *p*-value of 0.051 and 0.079, respectively, based on the McNemar test. The percentage of agreement between the MedCAT and SSNAP auditing methods when evaluated against the subsample reread is presented in Table S4.

TABLE 3 Table showing the performance metrics of SSNAP and MedCAT against the manually reviewed (ground truth) subset

	SSNAP, Cohen kappa (95%CI)	MedCAT, Cohen kappa (95% CI)	*McNemar p-value*
KCH			
AF	0.79 (0.65–0.93)	0.86 (0.75–0.97)	0.296
HTN	0.74 (0.60–0.88)	0.69 (0.53–0.84)	0.583
CCF	0.48 (0.21–0.76)	0.67 (0.45–0.89)	0.051
DM	0.87 (0.77–0.97)	0.90 (0.81–0.98)	0.369
GSTT			
AF	0.60 (0.41–0.80)	0.64 (0.48–0.80)	0.295
HTN	0.51 (0.33–0.70)	0.75 (0.60–0.91)	<0.001
CCF	0.51 (0.19–0.84)	0.64 (0.35–0.92)	0.079
DM	0.76 (0.63–0.89)	0.86 (0.75–0.96)	0.186

Note: The MedCAT model used a threshold heuristic of 5 and was applied to all available documents until discharge (2015-to-discharge). The level of agreement between the ground truth subsample reread and the auditing method was calculated using Cohen kappa, whereas McNemar test was applied to assess whether MedCAT classified patients more accurately than SSNAP (*p*-values are provided).

Abbreviations: AF, atrial fibrillation; CCF, congestive cardiac failure; DM, diabetes mellitus; GSTT, Guy's and St Thomas' Hospital; HTN, hypertension; KCH, King's College Hospital; SSNAP, Sentinel Stroke National Audit Programme.

DISCUSSION

The SSNAP is a national health care quality improvement programme supporting the delivery of evidence-based care for stroke. It has helped shape stroke services in the UK by measuring process of care, with data collected under time pressure in a continuous and contemporaneous manner. A hub-and-spoke system exists in the UK, where a hyperacute stroke unit (HASU) provides hyperacute intervention and care to a large area containing multiple smaller stroke units (SUs) that manage longer term rehabilitation needs. With a high annual incidence of stroke, the task of data collection is shared, with each hospital required to submit data for every patient that passes through its unit. In this project, we have examined the four comorbidities recorded by SSNAP over a 15-month period and evaluated the consistency of current audit practices, as well as a new machine learning-enabled method, MedCAT, against a manually reviewed set of patient episodes.

To evaluate an auditing method, a ground truth needs to be specified, and will inevitably require human operator involvement. Here, we have referred to our "subsample reread" as the ground truth and compared both the MedCAT and SSNAP methods against this dataset. Although this subsample reread is potentially vulnerable to the same risks of human error as in the SSNAP method, it was importantly not encumbered by the same time pressures that afflict SSNAP, focussed on comorbidities alone, and was performed completely by clinicians who would be better able to interpret the medical vernacular and extract the appropriate concepts. Moreover, the interrater consistency for both sites were strong, with a Cohen kappa greater than 0.80 for all comorbidities.

The KCH and PRUH sites operate both an HASU and an SU, with more than 1800 patient episodes annually between them. There was good consistency between SSNAP and our subsample reread for AF, hypertension, and diabetes. The F1 score for CCF was poor, primarily driven by low sensitivity. This is partly explained by the low prevalence of CCF within the subsample reread, with small numbers of detection errors incurring a large deterioration in performance. Importantly, the prevalence of heart failure in a nonenriched population aged more than 55 years is 7% [23], whereas the proportion of patients with CCF in the entire KCH-PRUH SSNAP data was 4.2%, therefore suggesting underrepresentation. As a clinical syndrome, diagnosis relies on the recognition of symptoms and signs, in combination with the interpretation of cardiac investigations to ensure accurate heart failure classification [24,25]. Consequently, diagnosing CCF from the clinical notes may be more challenging compared with the other comorbidities, with more junior clinical staff less likely to document the diagnosis explicitly or using poecilonyms like "heart failure" or "pulmonary oedema from heart failure." Although the diagnosis may be recorded in other documents, such as clinic letters, these may not be available to the acute team, especially if these documents originate from a different trust due to the organization of stroke services. In this scenario, data availability will rely on the recall of the patient and relative. A similar pattern is obtained from the data from GSTT, which only operates an SU. This operational

difference may explain the apparent deterioration in F1 score for AF, as the inpatient documentation is likely to have a greater emphasis on rehabilitation requirements rather than the aetiology of the stroke.

The low sensitivity of the SSNAP data compared with the subsample reread at KCH and GSTT would indicate sufficient information is present within the documents available at each site. The absence of time pressure, and requirement to extract only the comorbidities rather than all the SSNAP concepts, are likely to have contributed to greater sensitivity in the subsample reread, especially when a more comprehensive review of the records is required.

MedCAT uses NLP to extract concepts from the free text and maps them onto a standardized clinical vocabulary, SNOMED-CT. Intuitively, a more accurate picture of patients' stroke risk profiles will be obtained from a review of their entire medical histories. Clinical teams will review a patient's extensive history to identify potentially relevant stroke risk factors using sources from within and external to the hospital. This phenomenon is demonstrated in the performance of MedCAT, with higher F1 scores obtained when using all available documents (Table S2), although the effect appears less pronounced with the GSTT dataset. The machine learning-driven process of MedCAT achieved a higher F1 score for all comorbidities than the manually curated approach adopted by SSNAP. This increase in consistency is secondary to an improvement in sensitivity, with the superior performance of the machine learning-enabled approach reaching significance for CCF in the GSTT subsample reread. This therefore explains the lower F1 scores achieved by MedCAT when comparing against the SSNAP dataset, as it is simply reflecting the inherent errors of SSNAP.

There are several limitations to this audit. First, MedCAT was restricted to documents dated after January 2015, and did not include clinic letters or investigation reports, thereby reducing the sensitivity of the method. Modifying the documents available to MedCAT is possible and may improve sensitivity, particularly for "congestive cardiac failure" detection, especially if there are several closely associated but nonsynonymous concepts like "pulmonary oedema" and "heart failure." Second, the MedCAT model used to extract the comorbidities was initially trained on 500 manually annotated documents from KCH and PRUH electronic notes, thereby exposing the model to some of the data prior to this audit. The effect of this is likely to be small, as it represents 0.08% of the total number of documents. Importantly, there was no training performed on the GSTT data, with a similar level of performance achieved, highlighting the feasibility of using a validated general model across different sites. Further training of the model on local hospital documents could then be subsequently performed to optimize MedCAT to the local environment. Third, MedCAT requires the electronic medical notes to be centralized into a searchable state prior to execution. Although MedCAT and CogStack are both open source, setup of these systems can require significant upfront investment, depending on the size of the project, and may be considered a limiting factor to wider adoption. However, this needs to be considered against the running costs of a physician's time, and of a potential dedicated full-time curator. Fourth, despite the enriched population, the prevalence of the comorbidities is low, except for hypertension. Consequently, the training of the MedCAT model to the stroke-specific concepts was biased toward those concepts with greater representation within the population. It is unsurprising to see both AF and CCF perform worse than hypertension and diabetes. This is highlighted with errors where MedCAT misattributes the abbreviation "AF" to "atrial fibrillation" rather than "artificial feed" despite the nutritional context of the entry. Nevertheless, this issue can be addressed with further focussed training or more sophisticated NLP models (e.g., transformer-based models).

The UK SSNAP initiative requests participating centres to submit data within set deadlines to facilitate analysis and reporting. Each centre will utilize the existing clinical team, and may, as is the case at our three NHS sites, employ a dedicated trained operator to manually review the medical notes and extract the necessary concepts. This is representative of the arrangements across London, and likely throughout the rest of the United Kingdom. Our audit of stroke comorbidities collected by SSNAP shows reasonable performance of current methods, except for CCF, when compared with our subsample reread. The inconsistencies between the two datasets comprise a combination of factors, from limited physician involvement to time constraints in its collection. With the advancements in NLP, we have shown that machine learning-enabled data collection, in the form of MedCAT, has at least comparable concept extraction to the traditional manual processes used for SSNAP, and may potentially improve the quality of data when compared against our manual subsample.

MedCAT has several advantages. First, the speed of data collection is many orders of magnitude faster than manual review. With the CogStack ecosystem and MedCAT established within the hospital, patients' notes can be continuously monitored, providing near real-time data extraction if desired. Second, MedCAT is capable of interrogating huge volumes of data, applying the same level of scrutiny throughout the entire medical records, whether the patient has a short or extensive medical history. Third, deployment of different models is relatively simple. We have shown that a MedCAT model naïve to the local documents is noninferior to traditional methods employed by SSNAP. Additionally, it possesses the flexibility to adapt existing models through further training to address other medical and service questions. Moreover, deployment is primarily digitally delivered, potentially providing a more responsive and scalable system.

In conclusion, we have shown that machine learning-enabled data collection can support existing clinical and service initiatives, potentially improving the quality and speed of data extraction from existing clinical repositories. Clearly, extracting information located within densely packed resources is challenging for humans, especially medical notes, which are not structured, but are evolving documents, consisting of thoughts from multiple users over time. The scalability and flexibility of MedCAT therefore present an opportunity to revolutionize audit and research methods.

ACKNOWLEDGMENTS

The supporting infrastructure and code base received funding from NIHR Maudsley BRC, Health Data Research UK, UK Research and Innovation, London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare, Innovate UK, the NIHR Applied Research Collaboration South London, Office of Life Sciences (UK), and NHSX.

CONFLICT OF INTEREST

J.T. has received research grant funding support from Innovate UK, NHSX, Office of Life Sciences, Bristol-Meyers Squibb, and Pfizer; speaker honoraria from Bayer, Bristol-Meyers Squibb, Pfizer, and Goldman Sachs; hospitality from iRhythm Technologies; and copyright fees from Wiley-Blackwell; and owns public shares in Nvidia, Amazon, and Alphabet.

AUTHOR CONTRIBUTIONS

Anthony Shek: Data curation (equal), formal analysis (equal), writing–review & editing (equal). **Zhilin Jiang:** Data curation (equal), writing–review & editing (equal). **James Teo:** Data curation (equal), formal analysis (equal), supervision (equal), writing–review & editing (equal). **Joshua Au Yeung:** Data curation (equal), writing–review & editing (equal). **Ajay Bhalla:** Data curation (equal), writing–review & editing (equal). **Mark P. Richardson:** Supervision (equal), writing–review & editing (equal). **Yee Mah:** Conceptualization (equal), data curation (equal), formal analysis (equal), supervision (equal), writing–review & editing (equal).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

James Teo https://orcid.org/0000-0002-6899-8319
Yee Mah https://orcid.org/0000-0002-6652-9922

REFERENCES

1. Dutta D, Thornton D, Bowen E. Using population-based routinely collected data from the Sentinel Stroke National Audit Programme to investigate factors associated with discharge to care home after rehabilitation. *Clin Rehabil*. 2018;32(8):1108-1118.
2. Xu X-M, Vestesson E, Paley L, et al. The economic burden of stroke care in England, Wales and Northern Ireland: Using a national stroke register to estimate and report patient-level health economic outcomes in stroke. *Eur Stroke J*. 2018;3(1):82-91.
3. Great Britain, National Audit Office, Morse A. Department of Health: progress in improving stroke care. London: The Stationery Office; 2010.
4. Melamed RD, Khiabanian H, Rabadan R. Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. *BMC Bioinformatics*. 2014;15(Suppl 6):S3.
5. Bean DM, Kraljevic Z, Searle T, et al. Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. *Eur J Heart Fail*. 2020;22(6):967-974.
6. Mah Y-H, Nachev P, MacKinnon AD. Quantifying the impact of chronic ischemic injury on clinical outcomes in acute stroke with machine learning. *Front Neurol*. 2020;11:15. [cited 2020 May 9]. 10.3389/fneur.2020.00015/full
7. NHS England, Local Digital Roadmaps, Personalised Health and Care 2020. Using Data and Technology to Transform Outcomes for Patients and Citizens: A framework for Action [Internet]. Department of Health and Social Care; 2014. Available from: https://www.gov.uk/government/publications/personalised-health-and-care-2020/using-data-and-technology-to-transform-outcomes-for-patients-and-citizens
8. Electronic Health Records. Parlimentary Office of Science and Technology; 2016 Feb. Report No.: 519.
9. Office-based Physician Electronic Health Record Adoption [Internet]. Office of the National Coordinator for Health Information Technology - A division of the U.S. Department of Health and Human Services; 2019. Available from: dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php
10. WHO Global Observatory for eHealth, ProQuest (Firm). Global diffusion of eHealth: making universal health coverage achievable : report of the third global survey on eHealth [Internet]; 2016 [cited 2021 May 23]. Available from: https://ebookcentral.proquest.com/lib/qut/detail.action?docID=5910090
11. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*. 2008;15(1):36-39.
12. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical. in: proceedings of the 2nd clinical natural language processing workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019 [cited 2021 May 23]. p. 72-78. Available from: http://aclweb.org/anthology/W19-1909
13. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc*. 2019;26(11):1297-1304.
14. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. *J Biomed Inform*. 2020;109:103526.
15. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.
16. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artif Intell Med*. 2021;117:102083.
17. Royal College of Physicians, Clinical Effectiveness and Evaluation Unit on behalf of the Intercollegiate Stroke Working Part. Royal College of Physicians Sentinel Stroke National Audit Programme (SSNAP). National clinical audit annual results portfolio. March 2016-April 2017 [Internet]. Royal College of Physicians; 2017. Available from: http://bit.ly/1NHYlqH Last accessed October 24, 201
18. The quality of clinical coding in the NHS [Internet]. Capita; 2014. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/364476/The_quality_of_clinical_coding_in_the_NHS.pdf
19. Santos S, Murphy G, Baxter K, Robinson KM. Organisational factors affecting the quality of hospital clinical coding. *Health Inf Manag J*. 2008;37(1):25-37.
20. Alonso V, Santos JV, Pinto M, et al. Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. *J Med Syst*. 2020;44(3). [cited 2021 May 23] 10.1007/s10916-020-1532-x
21. National Data Guardian. Guidance: The Caldicott Principles [Internet]; 2020. Available from: https://www.gov.uk/government/publications/the-caldicott-principles
22. MATLAB Version 2020b. Matick, MA: The MathsWorks Inc; 2020. Computer Software.

Chapter 8: Summary and Conclusions

8.1 Summary of Key Findings and Contributions

This thesis aimed to explore the wealth of information in unstructured-text clinical documents, using scalable information search and retrieval tools to further our understanding of epilepsy. We have demonstrated the ability to i) accurately gather large retrospective cohorts, ii) sub-group patients based upon demographics, aetiology, and comorbidity, iii) and provide insights into healthcare service utilisation, treatment patterns, and patient health outcomes. Additionally, we have demonstrated the generalisability of these tools to other UK hospital sites and medical conditions.

Chapter 3 explored the first seizure patients' document trail from their presentation at the emergency department to their first specialist neurology follow-up, and ultimately to diagnosis. The most commonly reached diagnoses – after a first suspected seizure event – were epilepsy, cardiovascular disorders, and a single unknown episode. Our review of these records showed that in some cases the clinical pathway was not in keeping with NICE guidelines for the management of a first seizure event. This work was used to inform the development of neurology hot clinics at KCH.

Chapter 4 described the development of MedCAT, an accurate tool for extracting relevant clinical information from unstructured clinical records at scale. It uses a novel self-supervised machine learning algorithm for extracting clinical concepts using any standardised terminology. This approach overcomes many of the limitations of rule-based NLP techniques encountered in the previous chapter. Real-world validation of this tool demonstrated SNOMED-CT concept extraction from different EPR vendor systems at 3 large London hospitals with self-supervised training over 8.8B words from >17M clinical records and further fine-tuning with ~6,000 clinician annotated examples. We showed strong transferability (F-score: 0.94) between hospitals, datasets, and concept types, indicating cross-domain EPR-agnostic utility for accelerated clinical and research use cases.

Chapter 5 described the use of the CogStack digital infrastructure pipeline, including MedCAT, to identify all people with epilepsy across a 7-year period from neurology

outpatient clinics. We modified the SNOMED-CT terminology structure to enable the grouping of epilepsy-relevant SNOMED-CT concepts to reflect the multiple epilepsy subtypes. This allowed us to characterise the different epilepsy types across a large London hospital and described their rate of outpatient neurology clinic attendances and presence of comorbidities. We found that the majority of patients were classified under an unknown epilepsy type followed by focal, generalised, and lastly combined generalised and focal epilepsy.

Chapter 6 utilised the same epilepsy cohort as identified by the methodology stated in chapter 5. This chapter expanded upon the previous work by focusing on anti-seizure medication (ASM) prescription patterns, seizure freedom rates associated with the different ASM patterns, and explored correlations between different ASMs and combinations with a number of select idiopathic/associated symptoms. We found that levetiracetam was the most prescribed ASM monotherapy and a combination of Levetiracetam + Lamotrigine was the most prescribed polytherapy. Additionally, this study determined that polytherapy was associated with greater prevalence of idiopathic/associated symptoms than monotherapy.

Chapter 7 described the generalisability of the CogStack/MedCAT tools mentioned previously to applications beyond epilepsy and beyond a single NHS Trust. This chapter contains two examples of published work describing: 1) the extraction of cardiovascular related SNOMED-CT concepts documented in EPR records of patients admitted to hospital with a positive RT-PCR antigen test for SARS-CoV-2. This study found that pre-existing established cardiovascular disease (CVD) appeared to be a more important contributor to in-hospital mortality than cardiovascular risk factors without co-existent CVD, particularly in patients below the age of 70; 2) The cross-hospital site validation and extraction of routinely collected stroke comorbidities for the Sentinel Stroke National Audit Programme (SSNAP). This study demonstrated that machine learning-enabled data collection can support existing clinical and service initiatives and the results from the automated extraction, across different EPR providers and document types, were shown to be as accurate as current clinical data curation practises.

8.2 Practical Importance and Impact

8.2.1 Contributing to Local Organisation Insights and Evidence-based Guidelines

The tools explored in this thesis have the scope and ability to have real world benefits to patients by giving hospital stakeholders, including management and service staff, crucial real-time insights with the ultimate goal to monitor and improve care quality and service usage. This thesis has primarily explored these tools in the context of epilepsy. Epilepsy is a good use case to demonstrate the benefits of big data tools because it is a multifaceted neurological disorder which, whilst there is currently no cure for epilepsy, is widely considered a condition that can be successfully managed (Klein & Tyrlikova, 2020; Szaflarski et al., 2008). High quality care has been associated with improved seizure control and has shown to reduce the number of unnecessary future inpatient and emergency department follow-ups (Grinspan et al., 2018; Szaflarski et al., 2008; Wilner et al., 2014). Therefore, the value of utilising big data tools to establish effective care pathways for people with epilepsy is evident, as it has the potential to benefit both patients (improved outcomes in terms of disease control and QOL) and healthcare providers (reduction in both short and long-term resource costs). In a no-fee-for-service healthcare system such as the UK NHS, a healthy and happy patient is a cheap patient. Pathways are frequently audited to ensure a high standard of care. Although auditing care pathways is standard practice, the tools used in this thesis present the opportunity for audit automation, requiring less resources and empowering key stakeholders to extract more granular insights into patients' disease trajectory and hospital interactions, allowing them to better understand cross care pathways and key drivers of outcomes.

The National Institute for Health and Clinical Excellence (NICE) sets guidelines to standardise and optimise the provision of UK healthcare patient care pathways. Although NICE guidelines are evidence-based and intended to facilitate best practice, these guidelines themselves are not without their limitations. The primary data, which form the evidence for developing guidelines, are literature searches and selected expert testimonies (NICE, 2015). Susceptibility to bias relating to the nature of evidence, misconceptions, and personal recollections depending upon the beliefs of the developers, stakeholders and committee are some of the factors that may

confound the real-world validity and practical deployment of these guidelines to the local context (Z. Wang et al., 2018; Woolf et al., 1999). A further difficulty arises from the generalisation that such evidence is equally applicable to every individual, irrespective of variability in health urgency or circumstances (Franco et al., 2020).

In addition, NICE guidelines are often written for a single condition (e.g. epilepsy, diabetes mellitus, coronary heart disease, depression). However, patients often have several conditions at the same time and although there is guidance for multimorbidity management through NICE (NICE, 2016), this usually leads to the need to apply potentially conflicting recommendations from different care pathway guidelines in parallel. Aetiological factors and comorbidities such as anxiety, depression, or dementia present a high burden among epileptic patients. Therefore, many guidelines need to be interpreted and applied using clinical judgement, subject to specific settings and target groups (Z. Wang et al., 2018). For example, a clinician may need more knowledge than the rationale and evidence base provided by guidelines to choose the appropriate first line ASM for each patient with new onset epilepsy (Perrenoud & Novy, 2016). This is because guidelines are largely based on evidence from clinical trials, making them highly dependent on the study design and research questions of those clinical trials. In the case of ASM guidelines, the clinical trials they are based on focus mostly on the efficacy of first-line ASMs for new onset epilepsy rather than their profile of side effects; there are also currently no guidelines for the optimal order of polytherapy and drug switching (French, 2007).

Enterprise-level search and retrieval tools, which are accurate and validated, have the potential to gather a wide variety of data that can help supplement the guideline recommendations during clinical decision making. It can do this in two ways: 1) by feeding back real-patient hospital data into the evidence-base which informs these guidelines and 2) by highlighting the gaps in the guidelines and using aggregated historical patient data to inform care service providers how similar patients were managed in the past and help them make decision based on the respective outcomes. Ultimately, these tools should be implemented in a way that allows the gathering of insights into current practises which in turn will enable stakeholders to discuss, share, and eventually determine best clinical practises with the available resources.

Overall, guidelines combined with tools that provide an overview of who is using the system and why, should direct care pathways to be redesigned so that they form part of a larger integrated care system. The goal is to facilitate the early detection and identification of diseases and prompt intervention through appropriate management or treatment, thereby improving patients' lives and also potentially saving costs in the healthcare system.

8.2.2 Input of Structured vs Unstructured Textual Information

Epilepsy is a good example to demonstrate how standardising information contained in EPR systems could provide insights into the patterns of the heterogeneous nature of diseases and their management. Although large amounts of data are routinely collected from people with epilepsy, who often have frequent encounters with the healthcare system across a long period of time, the majority of descriptions around disease state and progression are contained within the unstructured text in EPR systems. This makes it challenging to aggregate and derive insights into disease and treatment outcomes at scale, due to clinicians' use of inconsistent terminology and epilepsy- and seizure-type descriptions, a problem that is further magnified by the recent changes in nomenclature from the old classification terminology.

One solution to the problem of unstructured data is to redesign the EPR system to allow the direct input of structured health-related information, through the selection of standardised terminology or classification codes, from a drop-down list. This can be incredibly useful in terms of reducing the requirement for data manipulation prior to analysing and gathering conclusions. However, recording a physician's day-to-day activities via unstructured text is unquestionably the most natural way of recording information, especially during patient communication. Limiting the input of EPR systems to standardised terminology and classification codes only works if diagnoses were always certain and already defined by medically relevant or related terms. However, this is not the case as over time our environment, knowledge, language and even diseases evolve. One could certainly argue that new structured terms could be added to a centralised terminology system but this would introduce complexity, delays, and redundancy in the terminology structure, and the risk to conflict with the

terminology used for previous patients, reducing the ability to compare patient data over time and thereby reducing the usefulness of the EPR system.

Therefore, keeping the more adaptable method of recording unstructured data, at least partially, is crucial. As our knowledge and medical classifications systems develop, they are incorporated into a widely accepted terminology. The advantage of computational tools such as NLP is that they can retrospectively infer structure from these data to provide the evidence for secondary layer actionable analytics. In reality, finding the correct balance between both structured and unstructured data is vital to capture both the most comprehensive details of a patient's health and their experience within a healthcare service, whilst maintaining a high degree of accurate details available in a format ready for rapid actionable insights.

8.3 End-to-End Open Science

Open science is a movement that strives to make the entire scientific research process as accessible and transparent as possible. This movement aims to increase both public and research collaboration and engagement through the dissemination of knowledge, and to improve the reproducibility of findings and transparent inspection of all tools/methodologies used.

Throughout this project, only open-source technologies were used to analyse routinely collected hospital data. Additionally, built from open-source python software packages, the tools created within this thesis are also available opensource[15] including tutorials on how to create your own MedCAT models and how use them[16].

Neither the NLP models used in this thesis, nor the source data can be made publicly available. This is due to the personally identifiable nature of the information found within the textual patient record dataset explored in this thesis.

The deidentification of the record is a commonly suggested solution “[PATIENT NAME]’s NHS number is [NHS NUMBER]” which can be done prior to training any

[15] https://github.com/CogStack/MedCAT/releases/tag/v0.1.9.9
[16] https://github.com/CogStack/working_with_cogstack/releases/tag/v1.0.0

language model. However non-obvious identifiers will remain, such as specific/unique occupations or descriptions for example “The prime minister of the UK contracted Covid-19” and thus this is considered de-identified and not anonymised data. Consequently, without auditing this information, de-identified data cannot be made publicly available.

The NLP models in this thesis are also not shared, because despite not purposefully containing identifiable information, they may be capable of leaking/exposing sensitive information. Several NLP model architectures have the potential to be reverse engineered to become generative models such as bidirectional LSTM (used in this thesis for meta-annotations) or Generative Pretrained Transformers (GPT). This may permit a training data extraction attack where the model re-generates data to which it has been exposed to during its training/finetuning (Carlini et al., 2021). This is because generative models attempt to produce plausible text based on statistical patterns learnt from their training data. For example, if the training data contains “Anthony Shek’s NHS number is 12345678”. The then-trained model when prompted with “Anthony Shek’s NHS number is …” has the possibility to re-generate and expose the sensitive information.

A risk assessment prior to the sharing of all language models should be produced and model sharing agreements can be set in place between trusted NHS organisations so that models are only to be used in an isolated air-gapped environment (NHS intranet only and not connected to the internet). This will minimise the risk of re-engineering models for nefarious means but will allow other NHS organisations to benefit from work done by other hospitals, whereby especially smaller organisations without the resources to build their own models would benefit.

8.4 Future Work and Considerations

8.4.1 Epilepsy Classification

Classification plays a central role in clinical epilepsy and epilepsy research. This thesis makes use of an enterprise-level data search and retrieval tool, and a new NLP approach to standardise clinical knowledge capture that maps to 2017 ILAE

classification of epilepsies. In addition, we demonstrate the capability of accurately gathering large retrospective cohorts and the capacity to subgroup patients based on patient-level demographic, clinical manifestation, and treatment features. Although this thesis has not explored how these extracted patient features and their patterns correspond to the current classification systems, future work presents the opportunity to investigate how disease patterns could be used to inform and potentially update classification systems.

An example of this is the grouping of people with epilepsy beyond epilepsy types, to include patterns of aetiology, comorbidities, demographics, and treatment response. This can potentially benefit the current 2017 ILAE classification of epilepsies as there are no clear criteria or guidelines on the role of these features towards classifying epilepsies (ILAE position paper).

NLP extraction methodologies have the potential to monitor the adoption and practical benefits of any changes to the nomenclature. Through providing a holistic picture of how patients' health information is recorded, it can provide further justification and a clear benefit to support changes to an existing classification system.

8.4.2 Standardised Terminologies

The standardised terminology system selected in this thesis to standardise extracted knowledge from unstructured EPR text is SNOMED-CT. Its use is important as it enables consistent information exchange, a feature that is fundamental to an interoperable EPR. It is a continuously developing terminology system which caters to a wide variety of users and use cases. SNOMED-CT already contains concept mapping to classification systems such as ICD-10 which is a system currently used in financial reimbursement processes, as a result automated clinical coding use cases are intrinsically possible despite training and validating NLP models for other purposes such as research or service evaluation.

SNOMED-CT has not been specifically developed for NLP use cases; consequently, this results in a number of limitations (discussed below). Therefore, the formation of a

standardised terminology with a specific design focus to assist NLP applications is favourable. This new terminology would become an intermediate step which would require mapping to a widely accepted terminology system, such as SNOMED-CT, to maintain interoperability.

As mentioned previously, in the UK, GPs already use SNOMED-CT codes as a structured form of reporting. However, as it is currently used it requires the creation of highly specific concepts which require the capture of contextual situations, for example: “family history of epilepsy”, “no family history of epilepsy”, “epilepsy clinic”, and “risk of epilepsy” are distinct concepts. Expanding the terminology in this way may not be feasible as it will in our example lead to duplicating the entire terminology multiple times only to capture each concepts’ context. In our methodology, we have used meta-annotations to replace the need for representing different contextual situations across different individual concepts. However, the general use and broad design of our current meta-annotations methodology may not be the solution for all use cases, for example where certain groups of concepts require specialised situational or locational extensions to their meaning but their references are not directly adjacent to the concept and are instead multiple sentences or paragraphs away. Alternative NLP relation extraction techniques with broader context spans to link different terms or concepts should be investigated as a possible alternative solution.

The concept extraction methodology (NER+L) used in this thesis contains a supervised annotation component where domain experts are required to manually annotate text and allows for the identification and creation of concepts which are not contained within SNOMED-CT. An example of this is a concept to represent “medication dose reduction (procedure)” which is common practice to increase the likelihood of capturing a seizure event during video telemetry. This knowledge can then be fed back to SNOMED-CT for other users to benefit from.

Another major limitation of SNOMED-CT is that the current hierarchical relationship structure is underdeveloped for several specialised domains including epilepsy. The current SNOMED-CT hierarchical structure of epilepsy – and seizure-related terms – does not reflect the current ILAE classifications systems. The SNOMED-CT terminology development team should consult and collaborate with specialist disease

classifications teams and curators who specialise in the field to develop a hierarchy structure which reflects the field's current releases of classification terminology and mapping to legacy terminology, whilst retaining the current use cases and benefits of SNOMED-CT. Ultimately, a careful well thought out process should be undertaken when adding new concepts and placing them into the correct locations in the terminology hierarchy; inconsiderate addition of terms into incorrect locations in the terminology system will diminish the value of the standardised terminology structure.

8.4.3 Considerations for Future Natural Language Processing Applications

When scaling ICT infrastructure, stakeholders need to be able to easily find the exact data they want to achieve their business objectives. The top priority when building an insights pipeline is incorporating efficient data flows that allow for data processing transparency. Users should be able to identify where in the pipeline outputs come from, and specifically from which input they are derived. Ideally this is achieved in a way that is readily accessible to identify data lineage or traceability to the source materials. Lastly, since AI processing is inherently a pipeline process, a sustainable architecture would be one where new NLP models can easily be validated, compared, and deployed. This is especially important because NLP is a rapidly advancing field with new technologies and model architectures being released, each one enabling better and more efficient performances for knowledge extraction tasks than the previous. Thereby, the ability to quickly compare new models to existing ones and – if warranted – interchange them could have a large impact on outputs and the insights derived from medical data.

8.4.4 Target Trial Emulation

This thesis demonstrated the ability to identify and describe demographic features and a variety of clinical and treatment patterns from routinely collected real-world data. However, no conclusions of causal relationships were made from these observational data.

In order to understand causal relationships, randomized controlled trials (RCT) are usually conducted and considered as "gold standard". However, due to the high cost

of these types of studies, both in terms of time and money, it is not always possible to conduct an RCT for every clinical intervention that is representative of the local clinical population. Therefore, many interventions still lack direct evidence and require clinicians to individually balance benefits and risks of each intervention, especially when patients' clinical features vary from the RCT's inclusion criteria or a combination of interventions is required. Each care decision based on individual patient circumstances is reflected in the routinely collected EPR data. One could therefore argue that the observational data captured in EPRs could be used as a proxy for when an RCT cannot be conducted, or its population sample was limited. In other words, if there is a question on comparative effectiveness or safety of an intervention which we would like to answer but cannot immediately conduct an RCT, observational data could be used to emulate a hypothetical randomised trial, allowing conclusions to be drawn. In this case, a hypothetical RCT without money or time constraints is our target of inference, known as the target trial.

Target trial emulation is not a new concept and has previously been demonstrated on quantifying individual treatment effect, i.e. the difference in outcome when a person is exposed to the treatment vs. not exposed to the treatment. A notable example of this is a study of interventions to prevent chronic migraine which included the ASMs valproate and topiramate (Stubberud et al., 2022). In the context of epilepsy, one could model the relative risk of seizure-free/not seizure-free outcomes between counter-factual ASM treatments.

Nevertheless, target trial emulations will be limited by the fact that the source data originates from daily practice which does not necessarily provide valid evidence for best practice (Meid et al., 2020). Using (real-world) data from routinely collected EPR for studies of comparative effectiveness can introduce many sources of bias, such as confounding, missing data, and misclassification, which future target trial emulation studies will have to navigate.

Chapter 9: Final Remarks

Epilepsy is highly heterogeneous in terms of its cause, demographics, clinical manifestations, treatment, and prognosis. The lives of most people with epilepsy continue to be adversely affected by gaps in knowledge, diagnosis, treatment, advocacy, education, legislation, and research. We are at a point in healthcare where care costs are high and yet come with no guarantee of a correspondingly good quality of care. We need to know whether people are at risk before they get sick. We need to know how to manage and slow disease progression of chronic conditions like epilepsy.

We are entering an era where large, complex organisations need to scale interactive computing with data to their entire organisation in a manner that is traceable, collaborative, secure, and human centred. There are so many factors which will affect one progression of health and until we can identify all the relevant information, we can only then decide which are feasible to collect accurately and are sufficient to provide valuable results. With the help of technology to sift through the mountains of unstructured data produced every day, information can be collated and processed faster to support changes to services or treatments in the context of limited resources. Supporting the shift to a healthcare system which can provide more value for its users. This thesis has described the application of an enterprise search and a new AI-based information retrieval tool and highlights its potential beneficial capabilities to promote a more comprehensive patient-centred care for epilepsy which can be generalised to diseases beyond.

References

Alexander, H. B., Broshek, D. K., & Quigg, M. (2018). Quality of life in adults with epilepsy is associated with anticonvulsant polypharmacy independent of seizure status. *Epilepsy & Behavior: E&B*, *78*, 96–99.

Amann, B., Grunze, H., Vieta, E., & Trimble, M. (2007). Antiepileptic drugs and mood stability. *Clinical EEG and Neuroscience: Official Journal of the EEG and Clinical Neuroscience Society* , *38*(2), 116–123.

American Psychiatric Association. (2013). In *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Medical Informatics*, *9*(7), e21929.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1409.0473

Barbour, K., Hesdorffer, D. C., Tian, N., Yozawitz, E. G., McGoldrick, P. E., Wolf, S., McDonough, T. L., Nelson, A., Loddenkemper, T., Basma, N., Johnson, S. B., & Grinspan, Z. M. (2019). Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. *Epilepsia*, *60*(6), 1209–1220.

Bean, D. M., Kraljevic, Z., Searle, T., Bendayan, R., Kevin, O. 'gallagher, Pickles, A., Folarin, A., Roguski, L., Noor, K., Shek, A., Zakeri, R., Shah, A. M., Teo, J. T. H., & Dobson, R. J. B. (2020). Angiotensin-converting enzyme inhibitors and

angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. In *European Journal of Heart Failure* (Vol. 22, Issue 6, pp. 967–974). https://doi.org/10.1002/ejhf.1924

Benbadis, S. R., Beniczky, S., Bertram, E., MacIver, S., & Moshé, S. L. (2020). The role of EEG in patients with suspected epilepsy. *Epileptic Disorders: International Epilepsy Journal with Videotape*, *22*(2), 143–155.

Ben-Menachem, E. (2002). Vagus-nerve stimulation for the treatment of epilepsy. *Lancet Neurology*, *1*(8), 477–482.

Berg, A. T. (2011). Epilepsy, cognition, and behavior: The clinical picture. *Epilepsia*, *52 Suppl 1*, 7–12.

Berg, A. T., Berkovic, S. F., Brodie, M. J., Buchhalter, J., Cross, J. H., van Emde Boas, W., Engel, J., French, J., Glauser, T. A., Mathern, G. W., Moshé, S. L., Nordli, D., Plouin, P., & Scheffer, I. E. (2010). Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia*, *51*(4), 676–685.

Berg, A. T., & Scheffer, I. E. (2011). New concepts in classification of the epilepsies: entering the 21st century. *Epilepsia*, *52*(6), 1058–1062.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–D270.

Bootsma, H. P., Ricker, L., Hekster, Y. A., Hulsman, J., Lambrechts, D., Majoie, M., Schellekens, A., de Krom, M., & Aldenkamp, A. P. (2009). The impact of side effects on long-term retention in three new antiepileptic drugs. *Seizure: The Journal of the British Epilepsy Association*, *18*(5), 327–331.

Brodie, M. J., & Kwan, P. (2012). Newer drugs for focal epilepsy in adults. In *BMJ* (Vol.

344, Issue jan26 2, pp. e345–e345). https://doi.org/10.1136/bmj.e345

Brodie, M. J., Richens, A., Yuen, A. W. C., & UK Lamotrigine/Carbamazepine Monotherapy Trial Group. (1995). Double-blind comparison of lamotrigine and carbamazepine in newly diagnosed epilepsy. *The Lancet*, *345*(8948), 476–479.

Campbell, E., Devenney, E., Morrow, J., Russell, A., Smithson, W. H., Parsons, L., Robertson, I., Irwin, B., Morrison, P. J., Hunt, S., & Craig, J. (2013). Recurrence risk of congenital malformations in infants exposed to antiepileptic drugs in utero. In *Epilepsia* (Vol. 54, Issue 1, pp. 165–171). https://doi.org/10.1111/epi.12001

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. and Oprea, A., 2021. Extracting training data from large language models. *In 30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).

Carlson, C., Dugan, P., Kirsch, H. E., & Friedman, D. (2014). Sex differences in seizure types and symptoms. In *Epilepsy & Behavior* (Vol. 41, pp. 103–108). https://doi.org/10.1016/j.yebeh.2014.09.051

Carr, E., Bendayan, R., Bean, D., Stammers, M., Wang, W., Zhang, H., Searle, T., Kraljevic, Z., Shek, A., Phan, H. T. T., Muruet, W., Gupta, R. K., Shinton, A. J., Wyatt, M., Shi, T., Zhang, X., Pickles, A., Stahl, D., Zakeri, R., … Dobson, R. J. B. (2021). Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *BMC Medicine*, *19*(1), 23.

Chong, M. S., & Smith, T. E. (2000). Anticonvulsants for the management of pain. In *Pain Reviews; London volume* (Vol. 7, Issue 3, pp. 119–149). Sage Publications Ltd. https://www.proquest.com/scholarly-journals/anticonvulsants-management-pain/docview/200001536/se-2

Chowdhury, F. A., Nashef, L., & Elwes, R. D. C. (2008). Misdiagnosis in epilepsy: a

review and recognition of diagnostic uncertainty. In *European Journal of Neurology* (Vol. 15, Issue 10, pp. 1034–1042). https://doi.org/10.1111/j.1468-1331.2008.02260.x

CogStack Documentation - Confluence. (2021, October 25). https://cogstack.atlassian.net/wiki/spaces/COGDOC/overview

Commission. (1964). A Proposed International Classification of Epileptic Seizures. *Epilepsia*, *5*(4), 297–306.

Commission. (1981). Proposal for revised clinical and electroencephalographic classification of epileptic seizures. Commission on classification and terminology of the ILAE. *Epilepsia*, *22*(4), 489–501.

Commission. (1989). Proposal for revised classification of epilepsies and epileptic syndromes. Commission on Classification and Terminology of the International League Against Epilepsy. *Epilepsia*, *30*(4), 389–399.

Dell'Orto, V. G., Belotti, E. A., Goeggel-Simonetti, B., Simonetti, G. D., Ramelli, G. P., Bianchetti, M. G., & Lava, S. A. G. (2014). Metabolic disturbances and renal stone promotion on treatment with topiramate: a systematic review. *British Journal of Clinical Pharmacology*, *77*(6), 958–964.

Department of Health. (2005). *Chief Executive's report to the NHS: December 2005*. https://webarchive.nationalarchives.gov.uk/20071204134909/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4124276

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805

Dickson, J. M., Jacques, R., Reuber, M., Hick, J., Campbell, M. J., Morley, R., &

Grünewald, R. A. (2018). Emergency hospital care for adults with suspected seizures in the NHS in England 2007–2013: a cross-sectional study. In *BMJ Open* (Vol. 8, Issue 10, p. e023352). https://doi.org/10.1136/bmjopen-2018-023352

Diener, H.-C., Tfelt-Hansen, P., Dahlöf, C., Láinez, M. J. A., Sandrini, G., Wang, S.-J., Neto, W., Vijapurkar, U., Doyle, A., Jacobs, D., & on behalf of the MIGR-003 StudyGroup. (2004). Topiramate in migraine prophylaxis. *Journal of Neurology*, *251*(8), 943–950.

Dixon, P. A., Kirkham, J. J., Marson, A. G., & Pearson, M. G. (2015). National Audit of Seizure management in Hospitals (NASH): results of the national audit of adult epilepsy in the UK. *BMJ Open*, *5*(3), e007325.

Duncan, J. S. (2019). Brain imaging in epilepsy. *Practical Neurology*, *19*(5), 438–443.

Electronic Medicines Compendium. (2021). *electronic medicines compendium (emc)*. Electronic Medicines Compendium (emc). https://www.medicines.org.uk/emc/

Falco-Walter, J. J., Scheffer, I. E., & Fisher, R. S. (2018). The new definition and classification of seizures and epilepsy. *Epilepsy Research*, *139*, 73–79.

Feldkamp, J., Becker, A., Witte, O., Scharff, D., & Scherbaum, W. (2012). Long-term anticonvulsant therapy leads to low bone mineral density — evidence for direct drug effects of phenytoin and carbamazepine on human osteoblast-like cells. In *Experimental and Clinical Endocrinology & Diabetes* (Vol. 108, Issue 01, pp. 37–43). https://doi.org/10.1055/s-0032-1329213

Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C.-S., Dykeman, J., Pringsheim, T., Lorenzetti, D. L., & Jetté, N. (2017). Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology*, *88*(3), 296–303.

Firkin, A. L., Marco, D. J. T., Saya, S., Newton, M. R., O'Brien, T. J., Berkovic, S. F.,

& McIntosh, A. M. (2015). Mind the gap: Multiple events and lengthy delays before presentation with a “first seizure.” *Epilepsia*, *56*(10), 1534–1541.

Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel, J., Jr, Forsgren, L., French, J. A., Glynn, M., Hesdorffer, D. C., Lee, B. I., Mathern, G. W., Moshe, S. L., Perucca, E., Scheffer, I. E., Tomson, T., Watanabe, M., & Wiebe, S. (2014). ILAE official report: a practical clinical definition of epilepsy. *Epilepsia*, *55*(4), 475–482.

Fisher, R. S., Cross, J. H., D’Souza, C., French, J. A., Haut, S. R., Higurashi, N., Hirsch, E., Jansen, F. E., Lagae, L., Moshé, S. L., Peltola, J., Roulet Perez, E., Scheffer, I. E., Schulze-Bonhage, A., Somerville, E., Sperling, M., Yacubian, E. M., & Zuberi, S. M. (2017). Instruction manual for the ILAE 2017 operational classification of seizure types. *Epilepsia*, *58*(4), 531–542.

Fisher, R. S., Cross, J. H., French, J. A., Higurashi, N., Hirsch, E., Jansen, F. E., Lagae, L., Moshé, S. L., Peltola, J., Roulet Perez, E., Scheffer, I. E., & Zuberi, S. M. (2017). Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 522–530.

Fonferko-Shadrach, B., Lacey, A. S., Roberts, A., Akbari, A., Thompson, S., Ford, D. V., Lyons, R. A., Rees, M. I., & Pickrell, W. O. (2019). Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. *BMJ Open*, *9*(4), e023232.

Franco, J. V. A., Arancibia, M., Meza, N., Madrid, E., & Kopitowski, K. (2020). Clinical practice guidelines: Concepts, limitations and challenges. *Medwave*, *20*(3), e7887.

French, J. A. (2007). Can evidence-based guidelines and clinical trials tell us how to treat patients? *Epilepsia*, *48*(7), 1264–1267.

Gastaut, H. (1969). Classification of the epilepsies. Proposal for an international classification. *Epilepsia*, *10*, Suppl:14–21.

Gastaut, H. (1970). Clinical and electroencephalographical classification of epileptic seizures. *Epilepsia*, *11*(1), 102–113.

GBD. (2019). Global, regional, and national burden of epilepsy, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology*, *18*(4), 357–375.

Gibbons, R. D., Hur, K., Brown, C. H., & Mann, J. J. (2009). Relationship between antiepileptic drugs and suicide attempts in patients with bipolar disorder. *Archives of General Psychiatry*, *66*(12), 1354–1360.

Gidal, B. E., Baltès, E., Otoul, C., & Perucca, E. (2005). Effect of levetiracetam on the pharmacokinetics of adjunctive antiepileptic drugs: a pooled analysis of data from randomized clinical trials. *Epilepsy Research*, *64*(1-2), 1–11.

Giussani, G., Bianchi, E., Beretta, S., Carone, D., DiFrancesco, J. C., Stabile, A., Zanchi, C., Pirovano, M., Trentini, C., Padovano, G., Colombo, M., Cereda, D., Tinti, L., Scanziani, S., Gasparini, S., Bogliun, G., Ferrarese, C., Beghi, E., & PRO-LONG Study Group. (2021). Comorbidities in patients with epilepsy: Frequency, mechanisms and effects on long-term outcome. *Epilepsia*, *62*(10), 2395–2404.

Golyala, A., & Kwan, P. (2017). Drug development for refractory epilepsy: The past 25 years and beyond. *Seizure: The Journal of the British Epilepsy Association*, *44*, 147–156.

Gray, M. (2017). Value based healthcare. *BMJ* , *356*, j437.

Grinspan, Z. M., Patel, A. D., Hafeez, B., Abramson, E. L., & Kern, L. M. (2018).

Predicting frequent emergency department use among children with epilepsy: A retrospective cohort study using electronic health data from 2 centers. *Epilepsia*, *59*(1), 155–169.

Guekht, A., Allen Hauser, W., Milchakova, L., Churillin, Y., Shpak, A., & Gusev, E. (2010). The epidemiology of epilepsy in the Russian Federation. In *Epilepsy Research* (Vol. 92, Issues 2-3, pp. 209–218). https://doi.org/10.1016/j.eplepsyres.2010.09.011

Hamid, H., Fodeh, S. J., Lizama, A. G., Czlapinski, R., Pugh, M. J., LaFrance, W. C., Jr, & Brandt, C. A. (2013). Validating a natural language processing tool to exclude psychogenic nonepileptic seizures in electronic medical record-based epilepsy research. *Epilepsy & Behavior: E&B*, *29*(3), 578–580.

Hippocrates. (2007). *On the Sacred Disease*. Library of Alexandria.

Ivanova, J. I., Birnbaum, H. G., Kidolezi, Y., Qiu, Y., Mallett, D., & Caleo, S. (2010). Economic burden of epilepsy among the privately insured in the US. *PharmacoEconomics*, *28*(8), 675–685.

Jackson, R., Kartoglu, I., Stringer, C., Gorrell, G., Roberts, A., Song, X., Wu, H., Agrawal, A., Lui, K., Groza, T., Lewsley, D., Northwood, D., Folarin, A., Stewart, R., & Dobson, R. (2018). CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. *BMC Medical Informatics and Decision Making*, *18*(1), 47.

Johannessen, S. I., & Landmark, C. J. (2010). Antiepileptic drug interactions - principles and clinical implications. *Current Neuropharmacology*, *8*(3), 254–267.

Keezer, M. R., Sisodiya, S. M., & Sander, J. W. (2016). Comorbidities of epilepsy: current concepts and future perspectives. *Lancet Neurology*, *15*(1), 106–115.

Ketter, T. A., Post, R. M., & Theodore, W. H. (1999). Positive and negative psychiatric effects of antiepileptic drugs in patients with seizure disorders. *Neurology*, *53*(5 Suppl 2), S53–S67.

Kingston, A., Robinson, L., Booth, H., Knapp, M., Jagger, C., & MODEM project. (2018). Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. *Age and Ageing*, *47*(3), 374–380.

Klein, P., & Tyrlikova, I. (2020). No prevention or cure of epilepsy as yet. In *Neuropharmacology* (Vol. 168, p. 107762). https://doi.org/10.1016/j.neuropharm.2019.107762

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., … Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, *47*(D1), D1018–D1027.

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A., Bendayan, R., Richardson, M. P., Stewart, R., Shah, A. D., Wong, W. K., Ibrahim, Z., Teo, J. T., & Dobson, R. J. B. (2021). Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, *117*, 102083.

Krumholz, A., Wiebe, S., Gronseth, G. S., Gloss, D. S., Sanchez, A. M., Kabir, A. A., Liferidge, A. T., Martello, J. P., Kanner, A. M., Shinnar, S., Hopp, J. L., & French, J. A. (2015). Evidence-based guideline: Management of an unprovoked first seizure in adults: Report of the Guideline Development Subcommittee of the

American Academy of Neurology and the American Epilepsy Society. *Neurology*, *84*(16). https://doi.org/10.1212/WNL.0000000000001487

Kwan, P., & Brodie, M. J. (2000). Early identification of refractory epilepsy. *The New England Journal of Medicine*, *342*(5), 314–319.

Kwan, P., & Brodie, M. J. (2006). Combination Therapy in Epilepsy. In *Drugs* (Vol. 66, Issue 14, pp. 1817–1829). https://doi.org/10.2165/00003495-200666140-00004

Landmark, C. J., Fossmark, H., Larsson, P. G., Rytter, E., & Johannessen, S. I. (2011). Prescription patterns of antiepileptic drugs in patients with epilepsy in a nation-wide population. *Epilepsy Research*, *95*(1-2), 51–59.

Lee, W. C., Arcona, S., Thomas, S. K., Wang, Q., Hoffmann, M. S., & Pashos, C. L. (2005). Effect of comorbidities on medical care use and cost among refractory patients with partial seizure disorder. *Epilepsy & Behavior: E&B*, *7*(1), 123–126.

Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: prescription for electronic drug information exchange. *IT Professional*, *7*(5), 17–23.

Llauradó, A., Santamarina, E., Fonseca, E., Olivé, M., Requena, M., Sueiras, M., Guzmán, L., Ballvé, A., Campos, D., Seijó, I., Abraira, L., Quintana, M., & Toledo, M. (2020). How soon should urgent EEG be performed following a first epileptic seizure? *Epilepsy & Behavior: E&B*, *111*. https://doi.org/10.1016/j.yebeh.2020.107315

Lyman, J. A., Scully, K., & Harrison, J. H., Jr. (2008). The development of health care data warehouses to support data mining. *Clinics in Laboratory Medicine*, *28*(1), 55–71, vi.

MacKenzie, S. L., Wyatt, M. C., Schuff, R., Tenenbaum, J. D., & Anderson, N. (2012). Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *Journal of the American Medical Informatics Association:*

JAMIA, *19*(e1), e119–e124.

Magiorkinis, E., Sidiropoulou, K., & Diamantis, A. (2010). Hallmarks in the history of epilepsy: epilepsy in antiquity. *Epilepsy & Behavior: E&B*, *17*(1), 103–108.

Majeed, A., Allwood, D., Foley, K., & Bindman, A. (2018). Healthcare outcomes and quality in the NHS: how do we compare and how might the NHS improve? *BMJ* , *362*, k3036.

Malmgren, K., Reuber, M., & Appleton, R. (2012). Differential diagnosis of epilepsy. *Oxford Textbook of Epilepsy and Epileptic Seizures*, 81–94.

Maloney, P., & LOINC Developers, L. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical*. https://academic.oup.com/clinchem/article-abstract/49/4/624/5641953

Marson, A., Burnside, G., Appleton, R., Smith, D., Leach, J. P., Sills, G., Tudur-Smith, C., Plumpton, C., Hughes, D. A., Williamson, P., Baker, G. A., Balabanova, S., Taylor, C., Brown, R., Hindley, D., Howell, S., Maguire, M., Mohanraj, R., & Smith, P. E. (2021a). The SANAD II study of the effectiveness and cost-effectiveness of levetiracetam, zonisamide, or lamotrigine for newly diagnosed focal epilepsy: an open-label, non-inferiority, multicentre, phase 4, randomised controlled trial. *Lancet*, *397*(10282). https://doi.org/10.1016/S0140-6736(21)00247-6

Marson, A., Burnside, G., Appleton, R., Smith, D., Leach, J. P., Sills, G., Tudur-Smith, C., Plumpton, C., Hughes, D. A., Williamson, P., Baker, G. A., Balabanova, S., Taylor, C., Brown, R., Hindley, D., Howell, S., Maguire, M., Mohanraj, R., & Smith, P. E. (2021b). The SANAD II study of the effectiveness and cost-effectiveness of valproate versus levetiracetam for newly diagnosed generalised and unclassifiable epilepsy: an open-label, non-inferiority, multicentre, phase 4, randomised controlled trial. *Lancet*, *397*(10282). https://doi.org/10.1016/S0140-

6736(21)00246-4

Marson, A. G., Al-Kharusi, A. M., Alwaidh, M., Appleton, R., Baker, G. A., Chadwick, D. W., Cramp, C., Cockerell, O. C., Cooper, P. N., Doughty, J., Eaton, B., Gamble, C., Goulding, P. J., Howell, S. J. L., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G. R., … Williamson, P. R. (2007a). The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *The Lancet*, *369*(9566), 1000.

Marson, A. G., Al-Kharusi, A. M., Alwaidh, M., Appleton, R., Baker, G. A., Chadwick, D. W., Cramp, C., Cockerell, O. C., Cooper, P. N., Doughty, J., Eaton, B., Gamble, C., Goulding, P. J., Howell, S. J. L., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G. R., … Williamson, P. R. (2007b). The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet*, *369*(9566), 1016.

Marson, A. G., Appleton, R., Baker, G. A., Chadwick, D. W., Doughty, J., Eaton, B., Gamble, C., Jacoby, A., Shackley, P., Smith, D. F., Tudur-Smith, C., Vanoli, A., & Williamson, P. R. (2007). A randomised controlled trial examining the longer-term outcomes of standard versus new antiepileptic drugs. The SANAD trial. *Health Technology Assessment* , *11*(37), iii – iv, ix – x, 1–134.

Mbizvo, G. K., Bennett, K., Simpson, C. R., Duncan, S. E., & Chin, R. F. M. (2018). Accuracy and utility of using administrative healthcare databases to identify people with epilepsy: a protocol for a systematic review and meta-analysis. *BMJ Open*, *8*(6), e020824.

Mbwana, J. S., Grinspan, Z. M., Bailey, R., Berl, M., Buchhalter, J., Bumbut, A.,

Danner, Z., Glauser, T., Glotstein, A., Goodkin, H., Jacobs, B., Jones, L., Kroner, B., Lapham, G., Loddenkemper, T., Maraganore, D. M., Nordli, D., & Gaillard, W. D. (2019). Using EHRs to advance epilepsy care. *Neurology. Clinical Practice*, *9*(1), 83–88.

McDonald, C. J. (1997). The barriers to electronic medical record systems and how to overcome them. *Journal of the American Medical Informatics Association: JAMIA*, *4*(3), 213–221.

Meinardi, H. (1999). International League against Epilepsy and its journal Epilepsia 1909--1999. *Hilversum, The Netherlands: Van Der Weij BV Grafische Bedrijven*.

Meid, A.D., Ruff, C., Wirbka, L., Stoll, F., Seidling, H.M., Groll, A. and Haefeli, W.E., 2020. Using the causal inference framework to support individualized drug treatment decisions based on observational healthcare data. *Clinical Epidemiology*, 12, p.1223.

Messenheimer, J., Ramsay, R. E., Willmore, L. J., Leroy, R. F., Zielinski, J. J., Mattson, R., Pellock, J. M., Valakas, A. M., Womble, G., & Risner, M. (1994). Lamotrigine therapy for partial seizures: a multicenter, placebo-controlled, double-blind, cross-over trial. *Epilepsia*, *35*(1), 113–121.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mizrahi, E. M. (1987). Neonatal seizures: problems in diagnosis and classification.

Epilepsia, *28 Suppl 1*, S46–S55.

Mizrahi, E. M., & Kellaway, P. (1987). Characterization and classification of neonatal seizures. *Neurology*, *37*(12), 1837–1844.

Mohanraj, R., & Brodie, M. J. (2005). Pharmacological outcomes in newly diagnosed epilepsy. *Epilepsy & Behavior: E&B*, *6*(3), 382–387.

Moran, N. F., Poole, K., Bell, G., Solomon, J., Kendall, S., McCarthy, M., McCormick, D., Nashef, L., Sander, J., & Shorvon, S. D. (2004). Epilepsy in the United Kingdom: seizure frequency and severity, anti-epileptic drug utilization and impact on life in 1652 people with epilepsy. *Seizure: The Journal of the British Epilepsy Association*, *13*(6), 425–433.

Muzina, D. J., Elhaj, O., Gajwani, P., Gao, K., & Calabrese, J. R. (2005). Lamotrigine and antiepileptic drugs as mood stabilizers in bipolar disorder. In *Acta Psychiatrica Scandinavica* (Vol. 111, Issue s426, pp. 21–28). https://doi.org/10.1111/j.1600-0447.2005.00523.x

Ngugi, A. K., Bottomley, C., Kleinschmidt, I., Sander, J. W., & Newton, C. R. (2010). Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. *Epilepsia*, *51*(5), 883–890.

NHS Connecting for Health, Great Britain: Department of Health, & Health and Social Care Information Centre. (2006). *OPCS Classifications of Interventions and Procedures*. The Stationery Office.

NHS Digital. (2006, December 15). *Outpatient data quality report 2003-04 and 2004-05*. NHS Digital. https://files.digital.nhs.uk/publicationimport/pub02xxx/pub02543/outp-qual.pdf

NHS England, 2021. *Plan, Do, Study, Act (PDSA) cycles and the model for improvement*.

NHSX. (2021, May 11). *The National Strategy for AI in Health and Social Care*. https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/

NICE. (2014, May 8). *Epilepsy | Treatment summary | BNF content published by NICE*. NICE: National Institute for Health and Care Excellence; NICE. https://bnf.nice.org.uk/treatment-summary/epilepsy.html

NICE. (2015). *Developing NICE Guidelines: The Manual*. National Institute for Health and Care Excellence (NICE).

NICE. (2016). *Multimorbidity: clinical assessment and management*. National Institute for Health and Care Excellence. https://www.nice.org.uk/guidance/ng56/resources/multimorbidity-clinical-assessment-and-management-pdf-1837516654789

NICE. (2021a, May 12). *NICE Guidance | Epilepsies: diagnosis and management*. NICE Guidance; NICE. https://www.nice.org.uk/guidance/cg137/chapter/1-Guidance#diagnosis-2

NICE. (2021b). *Epilepsies in children, young people and adults NICE guideline DRAFT* (NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE). https://www.nice.org.uk/guidance/gid-ng10112/documents/draft-guideline

Nicholas, J. M., Ridsdale, L., Richardson, M. P., Ashworth, M., & Gulliford, M. C. (2012). Trends in antiepileptic drug utilisation in UK primary care 1993-2008: cohort study using the General Practice Research Database. *Seizure: The Journal of the British Epilepsy Association*, *21*(6), 466–470.

Noachtar, S., & Rémi, J. (2009). The role of EEG in epilepsy: a critical review. *Epilepsy & Behavior: E&B*, *15*(1), 22–33.

O'Dwyer, M., Peklar, J., Mulryan, N., McCallion, P., McCarron, M., & Henman, M. C.

(2018). Prevalence and patterns of anti-epileptic medication prescribing in the treatment of epilepsy in older adults with intellectual disabilities. *Journal of Intellectual Disability Research: JIDR*, *62*(3), 245–261.

OECD. (2017). *Tackling wasteful spending on health*. OECD.

OpenSAFELY, *Emergency attendances (ECDS)- OpenSAFELY documentation.* Available at: https://docs.opensafely.org/data-sources/ecds/ [Accessed October 1, 2022].

Papanicolas, I., Woskie, L. R., & Jha, A. K. (2018). Health Care Spending in the United States and Other High-Income Countries. *JAMA: The Journal of the American Medical Association*, *319*(10), 1024–1039.

Patel, P., & Moshé, S. L. (2020). The evolution of the concepts of seizures and epilepsy: What's in a name? *Epilepsia Open*, *5*(1), 22–35.

Patsalos, P. N., Fröscher, W., Pisani, F., & van Rijn, C. M. (2002). The importance of drug interactions in epilepsy therapy. *Epilepsia*, *43*(4), 365–385.

Pavone, P., Polizzi, A., Marino, S. D., Corsello, G., Falsaperla, R., Marino, S., & Ruggieri, M. (2020). West syndrome: a comprehensive review. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, *41*(12), 3547–3562.

Peltola, J., Peltola, M., Raitanen, J., Keränen, T., Kharazmi, E., & Auvinen, A. (2008). Seizure-freedom with combination therapy in localization-related epilepsy. *Seizure: The Journal of the British Epilepsy Association*, *17*(3), 276–280.

Perrenoud, M. P., & Novy, J. (2016). First-Line Antiepileptic Drugs in Adults: From Guidelines to Personalized Medicine. *Epileptologie*, *33*, 232–239.

Personalised health and care 2020: a framework for action. (2021, October 27). https://www.gov.uk/government/publications/personalised-health-and-care-

2020/using-data-and-technology-to-transform-outcomes-for-patients-and-citizens

Perucca, E., French, J., & Bialer, M. (2007). Development of new antiepileptic drugs: challenges, incentives, and recent advances. *Lancet Neurology*, *6*(9), 793–804.

Perucca, P., & Gilliam, F. G. (2012). Adverse effects of antiepileptic drugs. *Lancet Neurology*, *11*(9), 792–802.

Petty, S. J., Paton, L. M., O'Brien, T. J., Makovey, J., Erbas, B., Sambrook, P., Berkovic, S. F., & Wark, J. D. (2005). Effect of antiepileptic medication on bone mineral measures. *Neurology*, *65*(9), 1358–1365.

Plevin, D., Jureidini, J., Howell, S., & Smith, N. (2018). Paediatric antiepileptic polytherapy: systematic review of efficacy and neurobehavioural effects and a tertiary centre experience. *Acta Paediatrica*, *2018/04/01*. https://doi.org/10.1111/apa.14343

Pressler, R. M., Cilio, M. R., Mizrahi, E. M., Moshé, S. L., Nunes, M. L., Plouin, P., Vanhatalo, S., Yozawitz, E., de Vries, L. S., Puthenveettil Vinayan, K., Triki, C. C., Wilmshurst, J. M., Yamamoto, H., & Zuberi, S. M. (2021). The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE Task Force on Neonatal Seizures. *Epilepsia*, *62*(3), 615–628.

Purcell, B., Gaitatzis, A., Sander, J. W., & Majeed, A. (2002). Epilepsy prevalence and prescribing patterns in England and Wales. *Health Statistics Quarterly / Office for National Statistics*, *15*(Autumn), 23–30.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & Others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Robertson, A., Cresswell, K., Takian, A., Petrakaki, D., Crowe, S., Cornford, T., Barber, N., Avery, A., Fernando, B., Jacklin, A., Prescott, R., Klecun, E., Paton, J., Lichtner, V., Quinn, C., Ali, M., Morrison, Z., Jani, Y., Waring, J., … Sheikh, A.

(2010). Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation. *BMJ* , *341*, c4564.

Roberts, S. J., Feely, M., & Bateman, D. N. (1998). Prescribing of anti-epileptic drugs in the northern and Yorkshire region: 1992–1995. *Seizure: The Journal of the British Epilepsy Association*, *7*(2), 127–132.

Rochon, P. A., & Gurwitz, J. H. (1997). Optimising drug treatment for elderly people: the prescribing cascade. *BMJ* , *315*(7115), 1096–1099.

Rodriguez Santana, I., Aragón, M. J., Rice, N., & Mason, A. R. (2020). Trends in and drivers of healthcare expenditure in the English NHS: a retrospective analysis. *Health Economics Review*, *10*(1), 20.

Rosenow, F., Schade-Brittinger, C., Burchardi, N., Bauer, S., Klein, K. M., Weber, Y., Lerche, H., Evers, S., Kovac, S., Hallmeyer-Elgner, S., Winkler, G., Springub, J., Niedhammer, M., Roth, E., Eisensehr, I., Berrouschot, J., Arnold, S., Schröder, M., Beige, A., … LaLiMo Study Group. (2012). The LaLiMo Trial: lamotrigine compared with levetiracetam in the initial 26 weeks of monotherapy for focal and generalised epilepsy--an open-label, prospective, randomised controlled multicenter study. *Journal of Neurology, Neurosurgery, and Psychiatry*, *83*(11), 1093–1098.

Rowan, A. J., Ramsay, R. E., Collins, J. F., Pryor, F., Boardman, K. D., Uthman, B. M., Spitz, M., Frederick, T., Towne, A., Carter, G. S., Marks, W., Felicetta, J., Tomyanovich, M. L., & VA Cooperative Study 428 Group. (2005). New onset geriatric epilepsy: a randomized study of gabapentin, lamotrigine, and carbamazepine. *Neurology*, *64*(11), 1868–1873.

Rutter, M. (1997). Comorbidity: concepts, claims and choices. *Criminal Behaviour and*

Mental Health: CBMH, *7*(4), 265–285.

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G. W., Moshé, S. L., Nordli, D. R., Perucca, E., Tomson, T., Wiebe, S., Zhang, Y.-H., & Zuberi, S. M. (2017). ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 512–521.

Searle, T., Kraljevic, Z., Bendayan, R., Bean, D., & Dobson, R. (2019). MedCATTrainer: A Biomedical Free Text Annotation Interface with Active Learning and Research Use Case Specific Customisation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. https://doi.org/10.18653/v1/d19-3024

Seidenberg, M., Pulsipher, D. T., & Hermann, B. (2009). Association of epilepsy and comorbid conditions. *Future Neurology*, *4*(5), 663–668.

Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Sheikh, A., Cornford, T., Barber, N., Avery, A., Takian, A., Lichtner, V., Petrakaki, D., Crowe, S., Marsden, K., Robertson, A., Morrison, Z., Klecun, E., Prescott, R., Quinn, C., Jani, Y., Ficociello, M., Voutsina, K., Paton, J., Fernando, B., … Cresswell, K. (2011). Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in “early adopter” hospitals. *BMJ* , *343*, d6054.

Silberstein, S. D., Lipton, R. B., Dodick, D. W., Freitag, F. G., Ramadan, N., Mathew, N., Brandes, J. L., Bigal, M., Saper, J., Ascher, S., Jordan, D. M., Greenberg, S.

J., Hulihan, J., & Topiramate Chronic Migraine Study Group. (2007). Efficacy and safety of topiramate for the treatment of chronic migraine: a randomized, double-blind, placebo-controlled trial. *Headache*, *47*(2), 170–180.

SNOMED CT Starter Guide. (2021, November 5). SNOMED Confluence. https://confluence.ihtsdotools.org/display/DOCSTART

Spencer, S. A., & Davies, M. P. (2012). Hospital episode statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. *BMJ Open*, *2*(6). https://doi.org/10.1136/bmjopen-2012-001651

Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 662–666.

Strozzi, I., Nolan, S. J., Sperling, M. R., Wingerchuk, D. M., & Sirven, J. (2015). Early versus late antiepileptic drug withdrawal for people with epilepsy in remission. *Cochrane Database of Systematic Reviews* , *2015*(2). https://doi.org/10.1002/14651858.CD001902.pub2

Stubberud, A., Gray, R., Tronvik, E., Matharu, M. and Nachev, P., 2022. Machine prescription for chronic migraine. *Brain communications*, 4(3), p.fcac059.

Szaflarski, J. P., Rackley, A. Y., Lindsell, C. J., Szaflarski, M., & Yates, S. L. (2008). Seizure control in patients with epilepsy: the physician vs. medication factors. *BMC Health Services Research*, *8*, 264.

Szczurkowska, P. J., Polonis, K., Becari, C., Hoffmann, M., Narkiewicz, K., & Chrostowska, M. (2021). Epilepsy and hypertension: The possible link for sudden unexpected death in epilepsy? *Cardiology Journal*, *28*(2), 330–335.

Taylor, R. S., Sander, J. W., Taylor, R. J., & Baker, G. A. (2011). Predictors of health-related quality of life and costs in adults with epilepsy: a systematic review.

Epilepsia, *52*(12), 2168–2180.

Thijs, R. D., Surges, R., O'Brien, T. J., & Sander, J. W. (2019). Epilepsy in adults. *The Lancet*, *393*(10172), 689–701.

Thorn, J. C., Turner, E., Hounsome, L., Walsh, E., Donovan, J. L., Verne, J., Neal, D. E., Hamdy, F. C., Martin, R. M., & Noble, S. M. (2016). Validation of the Hospital Episode Statistics Outpatient Dataset in England. *PharmacoEconomics*, *34*(2), 161–168.

Ungar, A., Ceccofiglio, A., Pescini, F., Mussi, C., Tava, G., Rafanelli, M., Langellotto, A., Marchionni, N., van Dijk, J. G., Galizia, G., Bonaduce, D., & Abete, P. (2017). Syncope and Epilepsy coexist in "possible" and "drug-resistant" epilepsy (Overlap between Epilepsy and Syncope Study - OESYS). *BMC Neurology*, *17*(1), 45.

Vajda, F. J. E., O'Brien, T. J., Lander, C. M., Graham, J., Roten, A., & Eadie, M. J. (2013). Teratogenesis in repeated pregnancies in antiepileptic drug-treated women. *Epilepsia*, *54*(1), 181–186.

Vega, D., Maalouf, N. M., & Sakhaee, K. (2007). Increased propensity for calcium phosphate kidney stones with topiramate use. *Expert Opinion on Drug Safety*, *6*(5), 547–557.

Verrotti, A., Tambucci, R., Di Francesco, L., Pavone, P., Iapadre, G., Altobelli, E., Matricardi, S., Farello, G., & Belcastro, V. (2020). The role of polytherapy in the management of epilepsy: suggestions for rational antiepileptic drug selection. *Expert Review of Neurotherapeutics*, *20*(2), 167–173.

Wang, Y. C., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13.

Wang, Y.-Q., Fang, Z.-X., Zhang, Y.-W., Xie, L.-L., & Jiang, L. (2020). Efficacy of the

ketogenic diet in patients with Dravet syndrome: A meta-analysis. *Seizure: The Journal of the British Epilepsy Association*, *81*, 36–42.

Wang, Z., Norris, S. L., & Bero, L. (2018). The advantages and limitations of guideline adaptation frameworks. *Implementation Science: IS*, *13*(1), 72.

Wang, H., Qin, K., Zakari, R., Lu, G., & Yin, J. (2022). Deep neural network-based relation extraction: an overview. *Neural Computing And Applications*, 34(6), 4781-4801.

Welcome to python.org. (2021, December 15). Python.org. https://www.python.org/

Werhahn, K. J., Trinka, E., Dobesberger, J., Unterberger, I., Baum, P., Deckert-Schmitz, M., Kniess, T., Schmitz, B., Bernedo, V., Ruckes, C., Ehrlich, A., & Krämer, G. (2015). A randomized, double-blind comparison of antiepileptic drug treatment in the elderly with new-onset focal epilepsy. *Epilepsia*, *56*(3), 450–459.

WHO. (2006). *Neurological disorders: public health challenges*. World Health Organization.

Williams, J. G., & Mann, R. Y. (2002). Hospital episode statistics: time for clinicians to get involved? *Clinical Medicine*, *2*(1), 34–37.

Wilner, A. N., Sharma, B. K., Thompson, A., Soucy, A., & Krueger, A. (2014). Diagnoses, procedures, drug utilization, comorbidities, and cost of health care for people with epilepsy in 2012. *Epilepsy & Behavior: E&B*, *41*, 83–90.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., & Others. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Woolf, S. H., Grol, R., Hutchinson, A., Eccles, M., & Grimshaw, J. (1999). Clinical

guidelines: Potential benefits, limitations, and harms of clinical guidelines. In *BMJ* (Vol. 318, Issue 7182, pp. 527–530). https://doi.org/10.1136/bmj.318.7182.527

World Health Organization. (2004). *The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 1: Tabular List*. World Health Organization.

Zaidi, A., Clough, P., Cooper, P., Scheepers, B., & Fitzpatrick, A. P. (2000). Misdiagnosis of epilepsy: many seizure-like attacks have a cardiovascular cause. *Journal of the American College of Cardiology*, *36*(1), 181–184.

Appendices

Supplementary Item 1: MedCATtrainer Guidelines

MedCATtrainer annotation guidelines

Version 1
January 2021

Introduction

What is MedCAT and MedCATtrainer?

MedCAT stands for MEDical Concept Annotation Tool. It is a simple information extraction and linking tool where documents containing unstructured/free text can be annotated/labelled with concepts from standardised clinical terminologies such as SNOMED CT. It has the ability to recognise and extract clinical concepts and the context in which they are mentioned.

MedCATtrainer is the user interface platform which allows users to interact with and input their domain expertise through annotating clinically valuable information held within unstructured text. These annotations are the training examples required to finetune MedCAT's algorithms ability to detect specialised clinical knowledge within and across different clinical document types. It uses active learning which means that as a user begins to annotate a set of documents using this tool, it will learn from previous annotations to assist in automatically annotating future documents with increasing accuracy. The user can then simply transition from annotating to validating each prediction. This enables users to annotate 100s of documents, at a rate much faster than by conventional methods, to create a large corpus of annotated documents.

This guide is divided into 3 sections:
1) MedCATtrainer User manual for annotators
2) Annotation guideline
3) MedCATtrainer project registration template

Section 1: MedCATtrainer User Interface

Getting started

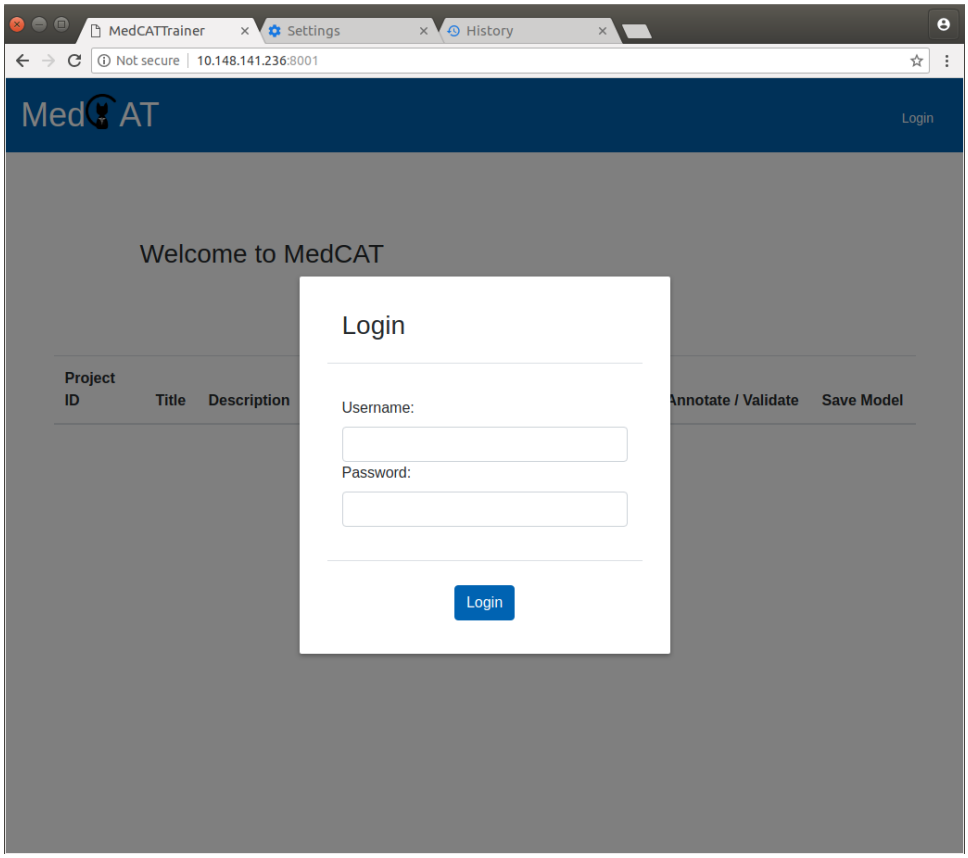
Before you begin you **must** have received the appropriate permissions from a MedCAT user with administrator permissions.

A MedCAT Admin user will send you 3 things:
Username
Password
A Link to access MedCAT from within the hospital intranet e.g. 10.39.240.51:8001.

1.1 Logging on

To access MedCAT within an NHS Trust, please follow this link and log in with your credentials:

Enter the link into the address bar of any web browser (Google Chrome, Internet Explorer, Firefox etc)
A user ID and password is required to log onto the web interface.



If you have any difficulties with logging in or wish to change your password, please contact the MedCAT team.

1.2 Logging off

In the top right-hand corner of every page there is a “logout” button. Please click here to log out.

1.3 Changing Username and Password

Only a user with MedCAT Admin Privileges can add users to a project and set usernames and passwords.

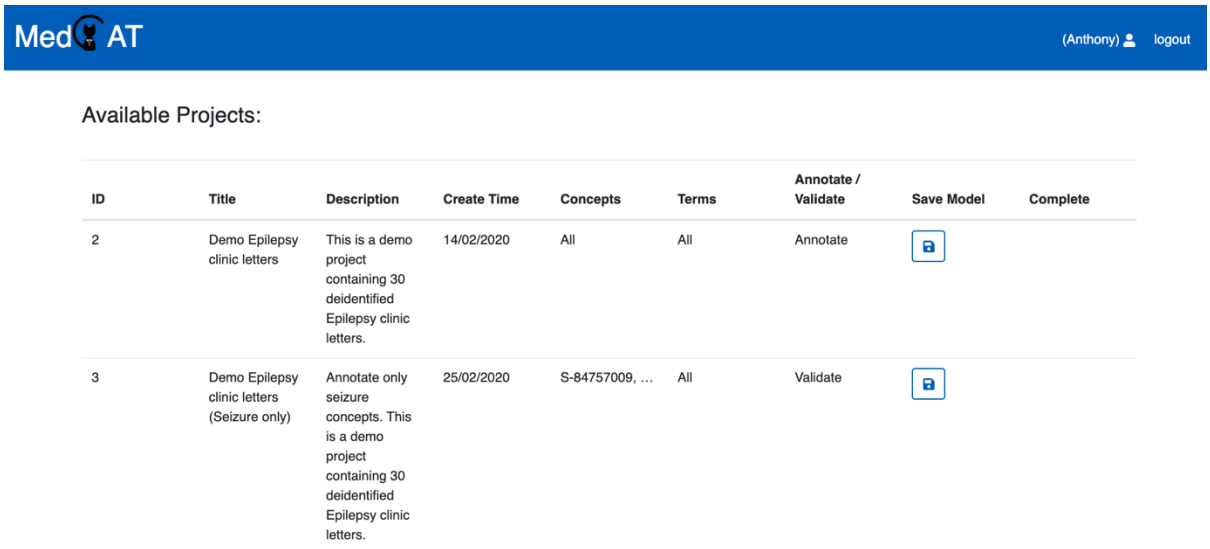
Username requirements:
150 characters or fewer. Letters, digits and @/./+/-/_ only.

Password requirements:
Not similar to your other personal information.
Contains at least 8 characters.
Not a commonly used password.
Not entirely Numeric.

Please contact your project manager or a person with MedCAT Admin Privileges to change your username or password.

MedCAT Dashboard

Once you have logged in. This will bring you to the main MedCAT Dashboard:

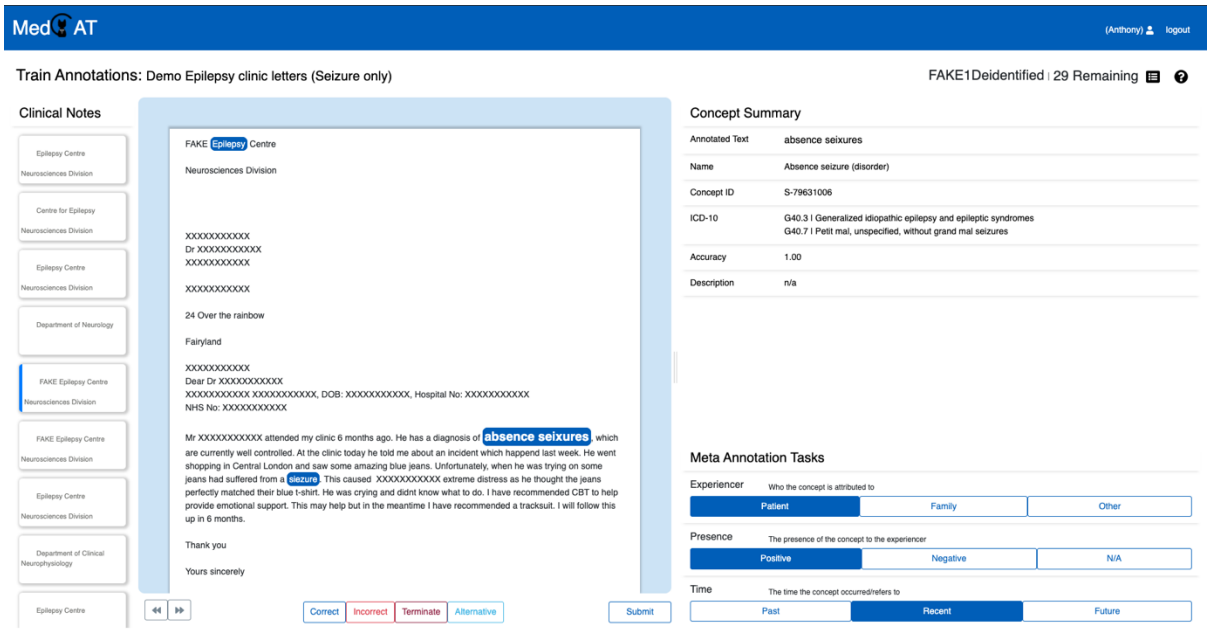


Project ID – The ID number of a project
Title – Project Title
Description – The Description of the project should contain information of the annotation task which the documents must be annotated by.
Create Time – Date which the project was created
Concepts – A positive Filter of the concepts which the documents are to be annotated.
Terms – A positive Filter of the Terms which the documents are to be annotated.
Annotate/Validate – Denotes whether these documents should be annotated/validated.
Save Model – To save the MedCAT annotation model in its current state.
Complete – If all documents are annotated. A green tick mark will appear here.

Annotation Interface

3.1 Annotation User Interface

Selecting an annotation project will bring you to the following interface:



3.2 Annotating a document

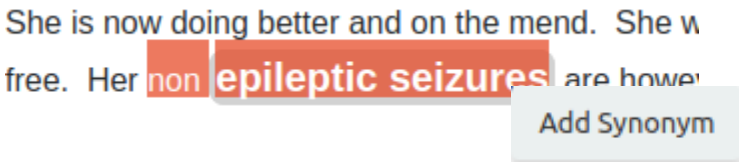
During the annotation of the document. The significance of each annotation within the document will be highlighted according to the colour scheme in the following table:

Annotation Label	Significance	Colour
Current concept	The current word/concept which the user is labelling	**BOLD**
Predicted Unlabelled concepts	These are words which have a predicted annotation label. These should be validated.	Grey
Correct labelled concepts	Word/concept with correct annotation label	Blue
Incorrect labelled concepts	Word/concept with an incorrect annotation label	Red
Terminate a concept	Word/concept which have been terminated from the dictionary due to their irrelevance.	Pink

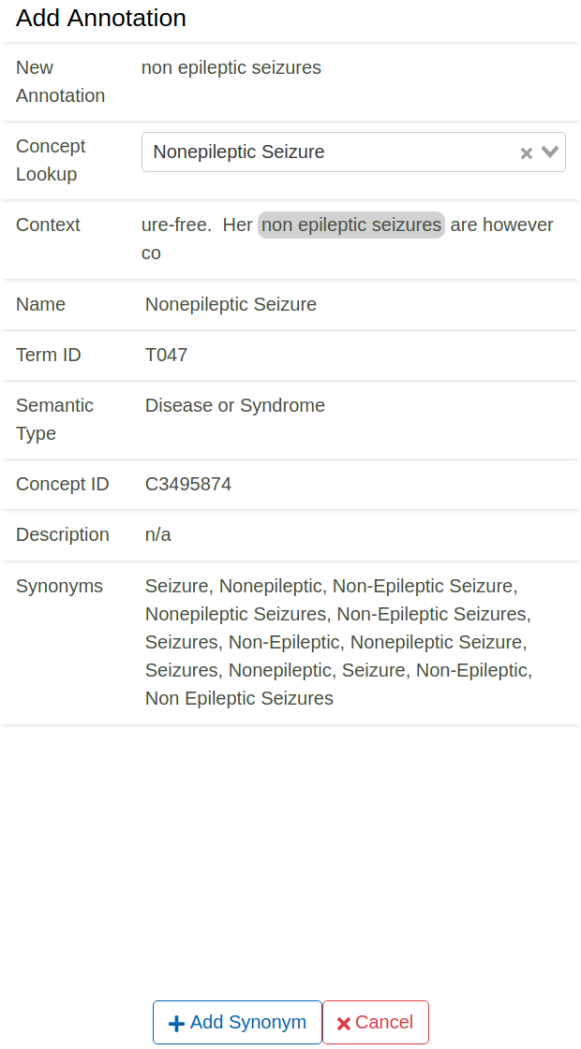
3.3 Adding a missing annotation

To add a missing annotation label. Highlight the section of text which you would label. Then right click this area. “Add Synonym” will appear.

In the below example: “epileptic seizures” has been annotated as epilepsy. However, this is incorrect. Instead “non epileptic seizures” should be annotated and provided with an annotation. The below image highlights the text of interest and right click to add a synonym.



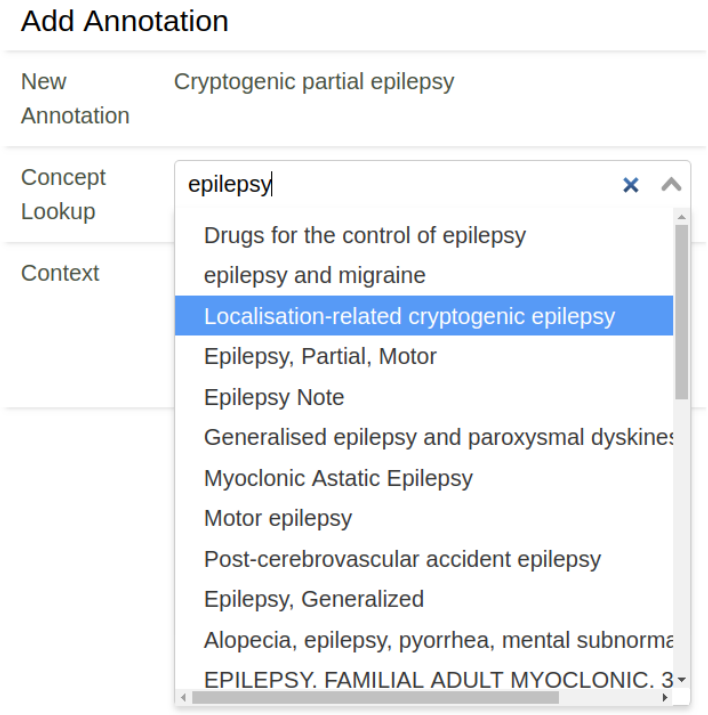
The user will be required to ensure that the New annotation has entirely covered the correct corresponding portion of text for annotation.



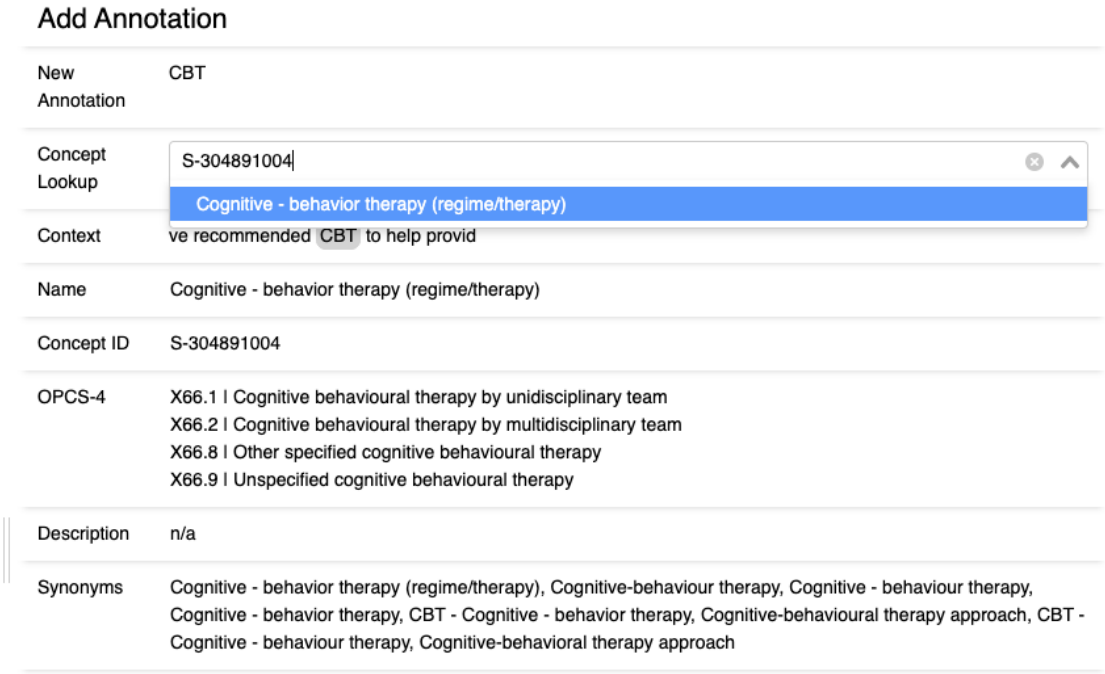
Click add synonym to add the concept to the document.

3.4 Search for a concept label

To attach an appropriate label to a concept: You can search SNOMED by concept name:



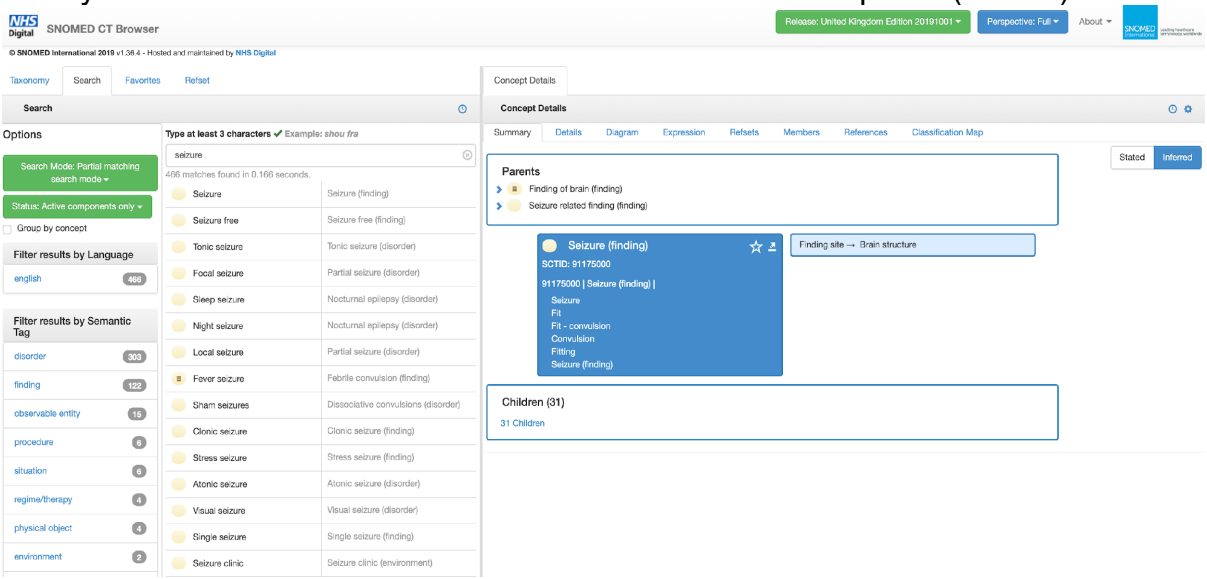
Or you can search for concept code using the prefix: S-<SCTID>:



External resources for searching for SNOMED concepts

If the project is using SNOMED as the concept database. You can search for further information of the here: https://termbrowser.nhs.uk/

Here you can search for terms and find the SNOMED Concept IDs (SCTID)



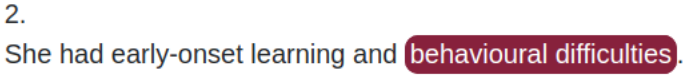
Note: Please check the version of SNOMED as these may differ. If you are unsure contact your project administrator.

3.5 Adding an alternative concept

Similar to adding a new concept you can add an alternative concept.
You will be required to add a synonym when a concept is labelled incorrectly given the context in which it is used. An example of this is “The patient with epilepsy had a fit” and “fit” is labelled as “healthy and well”. Instead “fit” should be labelled as “Seizure”.

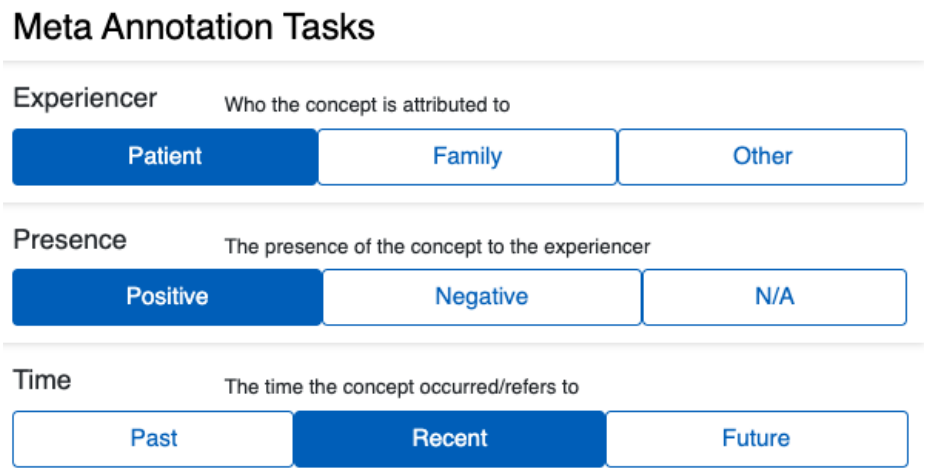
3.6 Terminating an annotation

If a word is being annotated completely incorrect to your objectives in your project. You can terminate the annotation. This will completely remove all annotations of that word from appearing in all future documents.



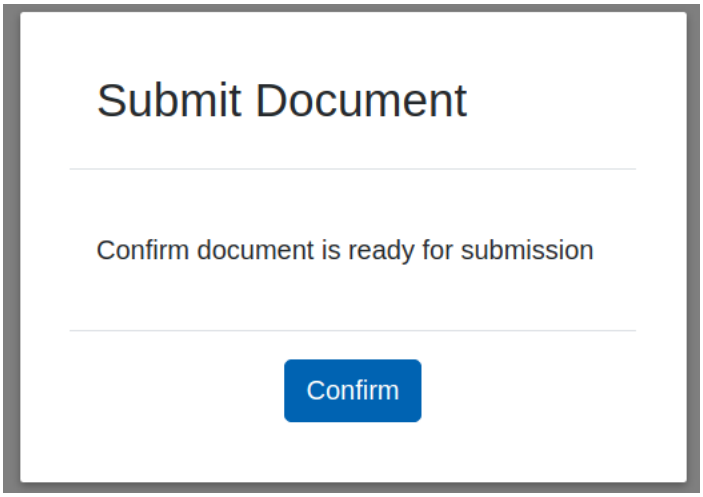
3.7 Add a Meta-annotation

If your project requires meta-annotations, they will have been set by your project administrator. A list of possible meta-annotations associated with the project will appear in the right-hand side box as shown below. Please choose the most appropriate meta-annotation for the annotation.



3.8 Submitting a document

Once the user is satisfied that all the required information contained within the document is correctly labelled. The User can then Submit the document by pressing the submit button at the bottom right hand side of the page. Users will then be provided with a summary page and required to confirm the submission to proceed. Once confirmed, this will automatically save the document which you have annotated and then move you onto the next document.



3.9 Shortcuts and Hotkeys

Selecting the grey question mark box (?) in the top right-hand corner will bring up this popup:

All Hotkeys are shown below:

Hotkey	Function
1	Correct annotation
2	Incorrect annotation
3	Terminate
4	Add synonym
Enter/Return (⏎)	Submit a document
Leftwards Arrow (←)	Previous annotation
Rightwards Arrow (→)	Next annotation
Up Arrow (↑)	Previous document
Down Arrow (↓)	Next document

Section 2: MedCAT project Annotation Guidelines

This Annotation guideline will walk through what is expected of a MedCATtrainer annotator, with an example project for annotating medical concepts. We encourage users to follow this template for future MedCAT projects in order to merge and build upon existing MedCAT models.

Annotation procedure

As the annotator you must read through each document very carefully to ensure that all concepts are labelled correctly in accordance with the task set above. Documents will be provided to you for annotation, one at a time. From which each annotation and meta annotation decision must be based solely upon the contextual information provided within that given document.

During the process of annotation, MedCAT may assist with some annotations. It is your job to ensure and validate that each annotation/meta-annotation MedCAT recommends within every document is correct to the best of your knowledge. This is extremely important, and your full attention is required at all times.

Annotation task

The annotation task is required to be defined prior to commencing each project. All annotators should familiarise themselves with the task to help to maintain consistency in their own work and between other annotators.

EXAMPLE annotation Task:
Your role as an annotator will be to annotate all mentions of 1) Disease and Symptoms with SCTID and 2) Medication with SCTID with the most correct labels to the best of your knowledge. During the annotation process, all annotations will require meta-annotations (contextual information of annotations). These meta annotations include: Experiencer Presence, and Temporality. In other words, as the annotator you must ask yourself throughout the annotation process and label: “Who is the subject/experiencing the highlighted concept?”, “Are they currently experiencing it?” and “When did it occur?”.

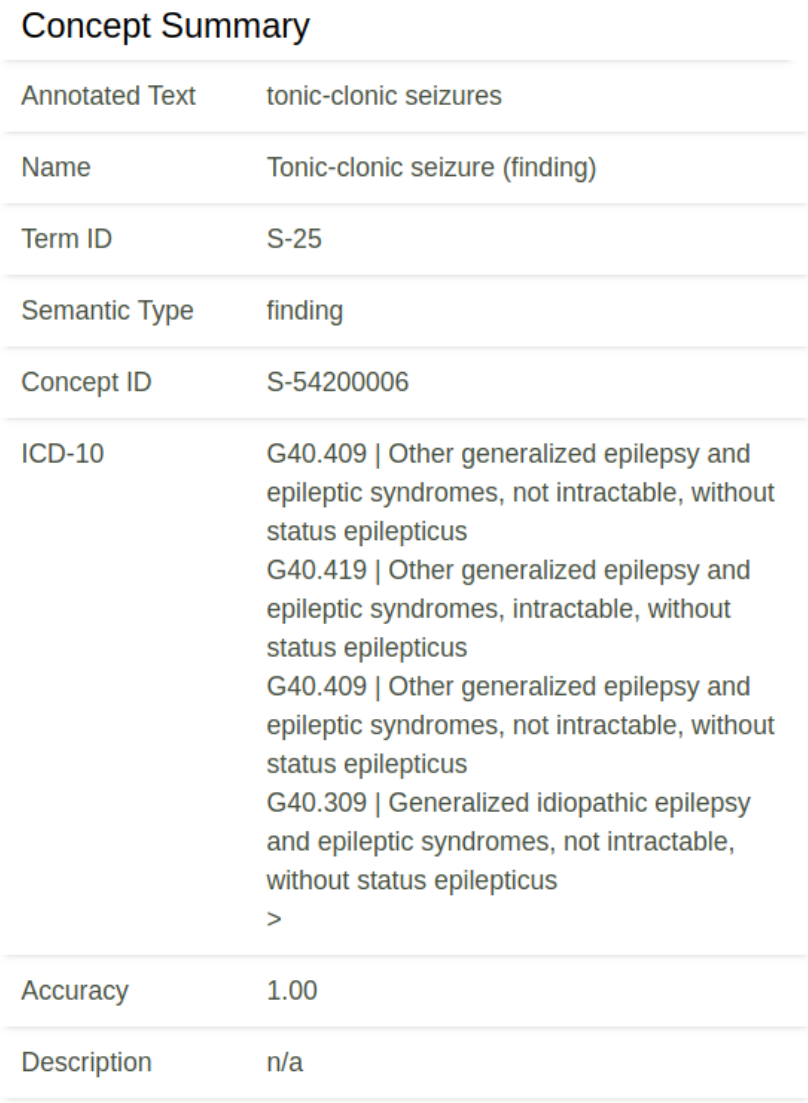
2.1 Standardised Clinical Terminology and Concept Filter

For each annotation project a Standardised clinical terminology is required to be selected. We recommend using SNOMED CT due to its comprehensiveness and its hierarchical terminology structure.

Additionally, we recommend limiting the number of concepts available to the annotators according to your projects use case. This will streamline the annotation process to only train concepts which are useful to your specific needs.

EXAMPLE:
As specified by the task, you will annotate all Disease, Symptoms and Drugs mentions with the corresponding unique SNOMED CT codes (SCTID). All SNOMED categories which are not previously list will not be available to annotate text.

The SCTID can be search for in MedCATtrainer through this format: prefix S-<code>
Or by their fully specified name (FSN): [example: Epilepsy (disorder)]



2.2 Adding annotations

There may be an instance where an annotation is unlabelled within a document. If this is the case, please follow instructions from Section 1: 3.3 Adding a missing annotation.
Keep in mind the format to enter a SCTID is: S-<code>, as shown in the above image.

2.3 Adding alternative concepts

If the annotation is labelled correctly but is incorrect within that given context. Please follow instructions from Section 1: 3.5 Adding an alternative concept.

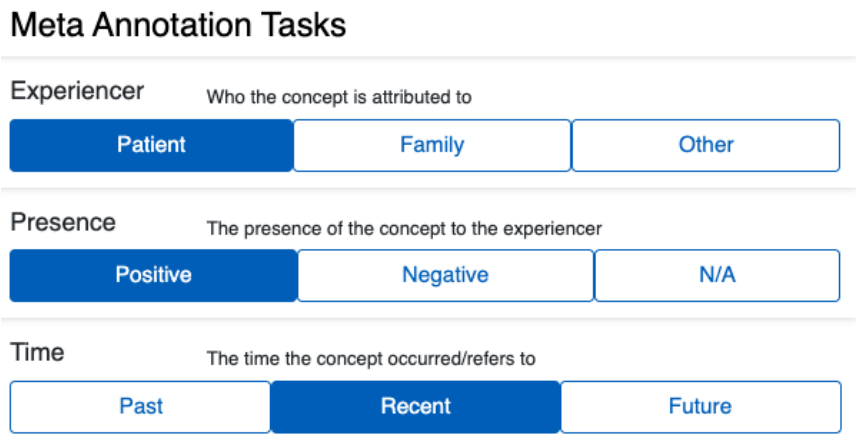
Meta-Annotations

This project will require the use of 3 meta-annotations:

1) Experiencer
2) Presence
3) Time

All "correct" identified concepts are required to have these meta-annotations to be completed. Meta-annotation options which are not selected will return a "not applicable" label to that annotation.

For instructions on how to label a meta-annotation within a document please follow the instructions from Section 1: 3.7 Add a Meta-annotation. Note: you will not be able to add meta-annotations which are not present from the below 3.



Note: The options highlighted in blue are the default values as they are the most common meta-annotation labels. If any of the meta-annotation section is left blank it will assume that the meta-annotation is "Not applicable" to that annotated concept.

3.1 Subject/Experiencer

The Experiencer task in this project is to identify the concepts reference to someone/something, or the subject of the conversation to which the concept relates.

- **Patient** – The concept is attributed to the patient.
  Example: The patient has been having seizures since he was 12.
- **Family** – The concept is attributed to the patient's family member.
  Example: The patient has a family history of epilepsy.
- **Other** – The concept is a noun used to identify any of a class of people, places, or things.
  Example: The Epilepsy nurse at the Epilepsy Clinic to specialise in epilepsy surgery. Here Epilepsy in all three annotations should be meta-annotated as other.

3.2 Presence

The Presence task in this project is to indicate the presence or existence of the concept. Do not confuse this with the sentiment of the sentence.

- **Positive** – The presence of the concept within the surrounding context.
  Example: The patient has non epileptic attacks.
- **Negative –** The absence of the concept within the surrounding context.
  Example: The patient no longer has seizures. He stopped taking his medication. The patient has not had a seizure today.
- **Hypothetical –** Concepts which are not present but are also not negatively mentioned but there is a possibility of occurrence.
  Example: If the patient develops depression, reschedule the appointment to an earlier date.
- **N/A** (no option selected) – If no option is selected the "Presence" meta-annotation will be treated as "N/A" for that concept.
  Example: "Epilepsy clinic". Here Epilepsy is correctly recognised as a disease but is the name of the hospital clinic.

3.3 Time/Temporality

The Temporality task in this project is to identify the appropriate time reference to which the concept relates to. The labels include:

- **Past** – A past event or historical presence of the concept.
  Example: PmHx seizures.
- **Recent** – Concept that occurs and relates to the reason for admission and/or current observations.
  Example: The patient collapsed at work last week**.**
  Example: The patient appears to be managing well with the current Keppra dose.
- **Future** – Events or presence of the concept which have the possibility to occur in the future.
  Example: If the patient continues to have seizures increase the dose to 250mg.

Example Annotation and Meta-annotation Table

The below table uses SNOMED-CT for annotation and the recommended meta-annotations options.

Text extract with highlighted annotation	Annotation	Meta-Annotation		
		Experiencer	Presence	Time
The patient had a fit during the clinic..	SCTID: S- 91175000 FSN: Seizure (finding)	Patient	True	Recent
The epilepsy nurse was...	SCTID: S-84757009 FSN: Epilepsy (disorder)	Other	N/A (unselected)	Recent
Your pt did not attend my Epilepsy clinic today	SCTID: S- 429279006 FSN: Seen in epilepsy clinic (finding)	Patient	False	Recent
Your pt did not attend my Epilepsy clinic today	SCTID: S-93401000000104 FSN: Did not attend epilepsy clinic (finding)	Patient	True	Recent
The pt is fit for discharge.	SCTID: S- 102499006 FSN: Fit and well (finding)	Patient	True	Recent
If patient suffers from additional sz increase AED dose	SCTID: S- 91175000 FSN: Seizure (finding)	Patient	Hypothetical	Future
The patient's sister also suffers from ASD.	SCTID: S- 35919005 FSN: Pervasive developmental disorder (disorder)	Family	True	Recent
I have some pain in my legs, maybe it is diabetic neuropathy.	SCTID: S- 230572002 FSN: Neuropathy due to diabetes mellitus (disorder)	Patient	Hypothetical	Recent
The pt's test this morning for CoVid-19 was positive.	SCTID: S-1240581000000104 FSN: Severe acute respiratory syndrome coronavirus 2 detected (finding)	Patient	True	Recent
PmHx: diabeties	SCTID: S- 73211009 FSN: Diabetes mellitus (disorder)	Patient	True	Past
There is no evidence suggestive of an AD diagnosis	SCTID: S- 26929004 FSN: Alzheimer's disease (disorder)	Patient	False	Recent
Diagnosis: ?first fit	SCTID: S- 91175000 FSN: Seizure (finding)	Patient	Hypothetical	Recent

Difficulties during annotating:

During the annotation process there may be words or phrases which will be tricky to annotate. In addition, searching for and selecting the most appropriate concept may be challenging at times. Try your best to annotate with the most correct concept in your opinion. Please contact the MedCAT team for any questions with the task. Whichever approach you decide to take it is important to keep in mind that consistency in the annotation process is crucial to producing a stable model.

5.1 Term dependencies

Consider the text:

"pt has simple and complex partial seizures"

How to proceed

- If there is a concept for which describes simple and complex partial seizures, then proceed to annotate with the concept.

- Otherwise only annotate partial seizures. Although the terms Simple and complex will be left out and the annotations will lose a degree of information. This will be addressed in the future.

5.2 Inappropriate meta-annotation availability

If at any point you feel that there is a meta-annotation required which is missing from the above 3 meta-annotations or is not compatible for your use case. Please contact your project administrator with any questions.

Please note that the recommended meta-annotations for most tasks are specified in the previous section. If you feel that the recommended Meta-annotations do not fit your project's use-case. MedCATtrainer has the ability to customise the type and options of meta-annotations. Different projects may require the use of different sets of meta-annotations dependent upon the project specific use case.

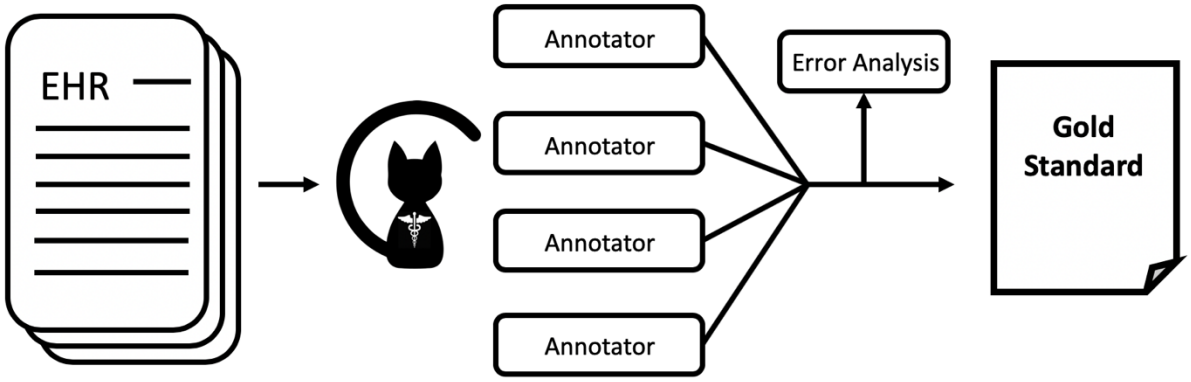
Outcome

Example:
At the end of the project. You will have annotated and meta-annotated all 1000 documents to the best of your knowledge. For all mentions of Disease and Symptoms and Drugs.

6.1 Optional: Inter-annotator agreement

Your results will then be cross validated against other annotators. Each and every discrepancy will then be discussed together until an agreement for each mismatching label is reached. Any discrepancies which cannot be resolved will be taken to an independent specialist reviewer.

Once a unanimous agreement of all annotations to correct labels has been reached. A “Gold standard” annotation set of documents would have been reached. This gold standard will then be used for NLP model development.



Definitions, Acronyms, Abbreviations

NLP – Natural Language Processing
EHR – Electronic Health Record, Electronic Patient Health Record (EPHR), Electronic patient record (EPR)

Annotation: A note by means of a comment added to a meaningful word from the sentence. For example, to annotate mentions of diseases with SCTID within a sentence: "The patient has no family history of epilepsy" {Epilepsy, SCTID: S-84757009}. This denotes that the underlined word is epilepsy within the SNOMED-CT under SCTID: S-84757009.

Meta-annotation: A meta-annotation is an annotation that can be applied to another pre-existing annotation. A meta-annotation is usually used to provide further information of the context of the pre-existing annotation. For example: "The patient has a negative family history of epilepsy" {Epilepsy, Experiencer: Family, Presence: Negative} This denotes that the annotation epilepsy is a negative mention and is mentioned in a context which refers to the Family.

SNOMED CT - Systematized Nomenclature of Medicine - Clinical Terms
SCTID – SNOMED concept ID
FSN – Fully Specified Name of SNOMED CT concepts. The opposite to SCTID.
Semantic Tags: Semantic tags are part of SNOMED descriptions. They are placed in parentheses at the end of FSNs when authoring concepts.

ICD – International Classification of Disease

Help and advice

If any questions arise or if any problems are encountered during the use of the program. Please contact the MedCAT team or your project administrator for any questions with the task.

The key to a successful annotation project is the consistency in annotation style/strategy by a single annotator and between different annotators.

Section 3: MedCATtrainer Project Description Template

The purpose of this project description template is to help standardise the annotation process across MedCAT projects, users and Healthcare Settings. We encourage users who require the MedCAT model for similar use cases to collaborate and build upon already existing NLP models through providing annotations of similar methodologies. Similar projects with compatible parameters and annotation methodologies will be able to be merged and/or continue pre-existing clinical NLP model development.

Project Title

The title of your MedCATtrainer annotation project.

Brief Project Summary

A brief description of the project and summary.

Aims and Objectives

The Aims and Objectives of the annotation project.

EXAMPLE:
The aim of this annotation project is to annotate all documents to the best of your knowledge for mentions of:

1) Disease and Symptoms with SCTID
2) Medication and drugs with SCTID

MedCAT users

List of all users and affiliations of those contributing to the project.

I. List of Administrators
II. List of Annotators

Defining the Data Set

This section should hold the following information:

- Name of Hospital trust/description of the dataset
- Date range of corpus of documents
- Number of documents annotated
- Medical Document type
- Any pre-processing steps of the free text documents
- Level of De Identification/redaction of information
- If available, to improve reproducibility/replication of research, instructions to recreate how the documents were retrieved and/or pre-processed. Note: If

medical documents were retrieved using CogStack or CRIS. Please link the search query or script.

EXAMPLE:
This dataset contains Neurology Clinic notes, discharge summaries and ITU discharge summaries.
From King's College Hospital NHS Foundation Trust uploaded between the datetime: 00:00:00 01/01/2013 to 00:00:00 01/10/2019.

Annotation methodology

Please specify in full any deviations to the recommended MedCAT annotation guidelines.

If different Meta-annotation options have been used. Please specify what were the options and any instruction/description for other annotators to follow.

Selected Standardised Clinical terminology

Please specify the standardised clinical terminology here: (UMLS, SNOMED CT, ICD-10, OPCS etc) the edition, and any other associated meta-data.

EXAMPLE:
This project uses SNOMED CT UK edition 01/04/2020 release with 28/03/2020 UK Drug extension available from NHS TRUD.

MedCAT concept Filter

Please provide a link to the selected Concepts of interest from the chosen standardised clinical terminology.
For large filters please attach a comprehensive list of all selected concepts to this document as a JSON file.

EXAMPLE:
For a simple annotation project which only requires the labelling of mentions of only the SNOMED concept "Epilepsy (disorder)" please put SCTID: S-84757009.

Project Timeline

Optional description of when the project occurred.

EXAMPLE:

Date	Checkpoint
4/11/2019	Introductory session for annotators on how to use MedCATtrainer
11/11/2019	End of pilot week
01/12/2019	1st check point to check progression
23/12/2019	Expected to complete Project 1. 2nd check point to check progression
01/02/2019	Completion both annotation projects

Outcome/Metrics/Model performance

This section should contain the following information:

- Did Double annotation occur? Inter-annotator agreement
- Performance of the NER+L model
- Evaluation/Performance of Meta-annotation models

Supplementary item 2: All extracted SNOMED CT Concepts

Table A – Symptom and Comorbidity Headings and Number of SCTID Child Concepts

Table B – All ASM Concepts and Number of SCTID Child Concepts

Table A) SNOMED_CT Epilepsy	Number of child SCTID concepts
Epilepsy	215
pregnant	41
breastfeeding mother	1
unrinary incontinence	39
intellectual disability	304
anxiety	276
depression	154
lethargy	2
headache	54
dizziness	64
rash	259
nausea	21
constipation	20
diarrhea	63
diabetes	356
hypertension	137
myocardial infarction	89
cerebrovasular accident	113
asthma	111
chronic obstructive lung disorder	40
fibromyalgia	2
dementia	128
liver failure	19
kidney disease	1290
seizure	164
dissociative convulsions	2
seizure free	2
Ketogenic diet	1
Vagal nerve stimulator in situ	1
DBS	3

Table B) SNOMED_CT ASM	Number of child SCTID concepts
Acetazolamide	140
Brivaracetam	29
Cannabidiol	39
Carbamazepine	141
Clobazam	153
Clonazepam	232
Eslicarbazepine	14
Ethosuximide	25
Everolimus	35
Gabapentin	233
Lacosamide	33
Lamotrigine	306
Levetiracetam	225
Oxcarbazepine	64
Perampanel	1
Phenobarbital	348
Phenytoin	106
Piracetam	12
Pregabalin	307
Primidone	45
Rufinamide	14
Stiripentol	11
Tiagabine	28
Topiramate	268
Sodium_valproate	153
Valproic_acid	207
Valproate_semisodium	18
Valproate	216
Vigabatrin	23
Zonisamide	82
Desmethylparamethadione	1
Potassium_bromide	1
Sulthiamine	10
Sodium_bromide	1
Felbamate	12
Dimethadione	1
Barbiturate_antiepileptic	8
Benzodiazepine	53
Oxazolidinedione	8
Succinimide	39
Paraldehyde	16
Beclamide	1
Clomethiazole	22
Caramiphen	4
Phenacemide	2
Magnesium_sulfate	2
Hydantoin_derivative_anticonvulsant	8
Retigabine	15
Pheneturide	1
Tiagabine_hydrochloride	2
Gabapentin_enacarbil	1
Valpromide	1

Supplementary Item 3: SNOMED Concept Group Extraction Performances

Supplementary Item 3A: SNOMED Disease Concept Group Extraction Performances

SNOMED group	Concept counts	FPs	FNs	TPs	P	R	F1
Epilepsy	604	16	75	520	0.970	0.874	0.920
Epilepsy type - Unknown	522	8	70	452	0.983	0.866	0.921
Seizure type - Unknown	522	8	70	452	0.983	0.866	0.921
Cerebrovascular accident	418	23	64	354	0.939	0.847	0.891
Hypertension	414	200	146	268	0.573	0.647	0.608
Chronic obstructive lung disorder	384	43	42	342	0.888	0.891	0.889
Diabetes	372	33	95	277	0.894	0.745	0.812
Seizure	253	25	54	199	0.888	0.787	0.834
Kidney disease	180	145	23	157	0.520	0.872	0.651
Dementia	68	4	2	66	0.943	0.971	0.957
Asthma	63	125	1	62	0.332	0.984	0.496
Myocardial infarction	43	3	2	41	0.932	0.953	0.943
Epilepsy type - Focal	36	4	2	34	0.895	0.944	0.919
Seizure type - Focal	36	4	2	34	0.895	0.944	0.919
Seizure type - Generalised	32	3	3	29	0.906	0.906	0.906
Anxiety	31	4	5	26	0.867	0.839	0.852
Depression	31	14	8	23	0.622	0.742	0.676
Epilepsy type - Generalised	26	3	3	23	0.885	0.885	0.885
Urinary incontinence	24	1	2	22	0.957	0.917	0.936
Diarrhoea	21	0	1	20	1.000	0.952	0.976
Epilepsy type - Combined generalised and focal	20	0	0	20	1.000	1.000	1.000
Vagal nerve stimulator in situ	17	2	0	17	0.895	1.000	0.944
Headache	17	1	0	17	0.944	1.000	0.971

Dissociative convulsions	16	0	0	16	1.000	1.000	1.000
Nausea	12	0	0	12	1.000	1.000	1.000
dizziness	11	0	0	11	1.000	1.000	1.000
Seizure free	11	4	1	10	0.714	0.909	0.800
Constipation	5	0	0	5	1.000	1.000	1.000
Rash	3	0	0	3	1.000	1.000	1.000
Lethargy	3	0	0	3	1.000	1.000	1.000
Pregnant	2	0	0	2	1.000	1.000	1.000
Ketogenic diet	1	0	0	1	1.000	1.000	1.000
DBS	1	0	0	1	1.000	1.000	1.000

Supplementary Item 3B: SNOMED anti-seizure medication (ASM) Concept Group Extraction Performances

SNOMED ASM group	**Concept Count**	**FPs**	**FNs**	**TPs**	**P**	**R**	**F1**
Carbamazepine	35	4	0	35	0.897	1.000	0.946
Lamotrigine	29	1	0	29	0.967	1.000	0.983
Valproate	20	0	0	20	1.000	1.000	1.000
Clobazam	20	3	0	20	0.870	1.000	0.930
Levetiracetam	20	4	0	20	0.833	1.000	0.909
Topiramate	15	0	0	15	1.000	1.000	1.000
Rufinamide	3	0	0	3	1.000	1.000	1.000
Phenytoin	3	0	0	3	1.000	1.000	1.000
Zonisamide	2	0	0	2	1.000	1.000	1.000
Oxcarbazepine	2	0	0	2	1.000	1.000	1.000
Pregabalin	2	1	0	2	0.667	1.000	0.800
Perampanel	1	0	0	1	1.000	1.000	1.000
Phenobarbital	1	0	0	1	1.000	1.000	1.000

Lacosamide	1	0	0	1	1.000	1.000	1.000
Gabapentin	1	0	0	1	1.000	1.000	1.000
Brivaracetam	0	0	0	0	0.000	0.000	0.000
Piracetam	0	0	0	0	0.000	0.000	0.000
Everolimus	0	0	0	0	0.000	0.000	0.000
Primidone	0	0	0	0	0.000	0.000	0.000
Ethosuximide	0	0	0	0	0.000	0.000	0.000
Stiripentol	0	0	0	0	0.000	0.000	0.000
Tiagabine	0	0	0	0	0.000	0.000	0.000
Eslicarbazepine	0	0	0	0	0.000	0.000	0.000
Clonazepam	0	0	0	0	0.000	0.000	0.000
Cannabidiol	0	0	0	0	0.000	0.000	0.000
Vigabatrin	0	0	0	0	0.000	0.000	0.000
Acetazolamide	0	0	0	0	0.000	0.000	0.000

Supplementary Item 4: Reclassification of SNOMED Epilepsy Concepts

Unknown	Focal	Generalised	Combined generalised and focal	None
Epilepsy in mother complicating pregnancy (disorder)	Refractory occipital lobe epilepsy (disorder)	Developmental delay, epilepsy, neonatal diabetes syndrome (disorder)	Lennox-Gastaut syndrome (disorder)	Benign non-familial neonatal convulsions (disorder)
Epilepsy in mother complicating childbirth (disorder)	Refractory frontal lobe epilepsy (disorder)	Atypical absence epilepsy (disorder)	Cryptogenic Lennox-Gastaut syndrome (disorder)	
Epilepsy, not refractory (disorder)	Epilepsy due to infectious disease of central nervous system (disorder)	Generalized non-convulsive epilepsy (disorder)	Symptomatic Lennox-Gastaut syndrome (disorder)	
Refractory epilepsy (disorder)	Epilepsy due to perinatal stroke (disorder)	Benign myoclonic epilepsy in infancy (disorder)	Severe myoclonic epilepsy in infancy (disorder)	
Epilepsy (disorder)	Epilepsy due to cerebrovascular accident (disorder)	Generalized epilepsy (disorder)	Generalized epilepsy with febrile seizures plus (disorder)	
Epilepsy undetermined whether focal or generalized (disorder)	Epilepsy co-occurrent and due to degenerative brain disorder (disorder)	Benign neonatal familial convulsions (disorder)	Benign adult familial myoclonic epilepsy (disorder)	
Epilepsy with continuous spike wave during slow-wave sleep (disorder)	Epilepsy due to immune disorder (disorder)	Myoclonic epilepsy of early childhood (disorder)		
Nocturnal epilepsy (disorder)	Epilepsy co-occurrent and due to demyelinating disorder (disorder)	Juvenile absence epilepsy (disorder)		
Female restricted epilepsy with intellectual disability syndrome (disorder)	Epilepsy co-occurrent and due to dementia (disorder)	Epilepsy with grand mal seizures on awakening (disorder)		
Pyridoxal 5-phosphate dependent epilepsy (disorder)	Epileptic dementia with behavioral disturbance (disorder)	Unverricht-Lundborg syndrome (disorder)		
Epilepsy telangiectasia syndrome (disorder)	Localization-related(focal)(partial)idiopathic epilepsy and epileptic syndromes with seizures of localized onset (disorder)	Lafora disease (disorder)		
Symptomatic generalized epilepsy (disorder)	Benign frontal epilepsy of childhood (disorder)	Myoclonic epilepsy with ragged red fibers (disorder)		
Tonic-clonic epilepsy (disorder)	Benign psychomotor epilepsy of childhood (disorder)	Cryptogenic myoclonic epilepsy (disorder)		

Drug-induced epilepsy (disorder)	Benign atypical partial epilepsy in childhood (disorder)	Early infantile epileptic encephalopathy with suppression bursts (disorder)		
Narcotic withdrawal epilepsy (disorder)	Childhood epilepsy with occipital paroxysms (disorder)	Progressive myoclonic epilepsy (disorder)		
Menstrual epilepsy (disorder)	Benign occipital epilepsy of childhood - early onset variant (disorder)	Refractory generalized convulsive epilepsy (disorder)		
Alcohol-induced epilepsy (disorder)	Benign occipital epilepsy of childhood - late onset variant (disorder)	Idiopathic generalized epilepsy (disorder)		
Stress-induced epilepsy (disorder)	Primary inherited reading epilepsy (disorder)	Benign neonatal convulsions (disorder)		
Febrile infection related epilepsy syndrome (disorder)	Amygdalo-hippocampal epilepsy (disorder)	Retropulsion petit mal (disorder)		
Photogenic epilepsy (disorder)	Rhinencephalic epilepsy (disorder)	Refractory generalized nonconvulsive epilepsy (disorder)		
Centrencephalic epilepsy (disorder)	Frontal lobe epilepsy (disorder)	Primary generalized absence epilepsy (disorder)		
Reflex epilepsy (disorder)	Cingulate epilepsy (disorder)	Childhood absence epilepsy (disorder)		
Myoclonic seizure (disorder)	Anterior frontopolar epilepsy (disorder)	Juvenile myoclonic epilepsy (disorder)		
Epileptic encephalopathy (disorder)	Orbitofrontal epilepsy (disorder)	Generalized convulsive epilepsy (disorder)		
Syntaxin binding protein 1 encephalopathy with epilepsy (disorder)	Dorsolateral epilepsy (disorder)	Generalized non-convulsive absence epilepsy (disorder)		
Acute encephalopathy with biphasic seizures and late reduced diffusion (disorder)	Opercular epilepsy (disorder)	Refractory juvenile myoclonic epilepsy (disorder)		
Infantile epileptic dyskinetic encephalopathy (disorder)	Non-progressive Kozhevnikow syndrome (disorder)	Refractory myoclonic epilepsy (disorder)		
Early-onset epileptic encephalopathy, cortical blindness, intellectual disability, facial dysmorphism syndrome (disorder)	Parietal lobe epilepsy (disorder)	Progressive myoclonus epilepsy with ataxia (disorder)		

Potassium voltage-gated channel subfamily Q member 2 related epileptic encephalopathy (disorder)	Occipital lobe epilepsy (disorder)	Refractory idiopathic generalized epilepsy (disorder)		
Myoclonic epilepsy in non-progressive encephalopathy (disorder)	Chronic progressive epilepsia partialis continua of childhood (disorder)	Generalized epilepsy and paroxysmal dyskinesia syndrome (disorder)		
Pyridoxine-dependent epilepsy (disorder)	Localization-related cryptogenic epilepsy (disorder)	Jeavons syndrome (disorder)		
	Localization-related idiopathic epilepsy (disorder)	Early-onset Lafora body disease (disorder)		
	Intractable idiopathic partial epilepsy (disorder)	Progressive myoclonic epilepsy with dystonia (disorder)		
	Chronic progressive epilepsia partialis continua (disorder)	Action myoclonus renal failure syndrome (disorder)		
	Benign childhood epilepsy with centrotemporal spikes, refractory (disorder)	Cryptogenic late-onset epileptic spasms (disorder)		
	Benign childhood epilepsy with centrotemporal spikes, non-refractory (disorder)	Progressive myoclonic epilepsy type 5 (disorder)		
	Benign Rolandic epilepsy (disorder)	Progressive myoclonic epilepsy type 6 (disorder)		
	Rolandic epilepsy, speech dyspraxia syndrome (disorder)	Progressive myoclonic epilepsy type 3 (disorder)		
	Benign occipital lobe epilepsy (disorder)	Progressive myoclonic epilepsy type 8 (disorder)		
	Intractable simple partial epilepsy (disorder)	Familial infantile myoclonic epilepsy (disorder)		
	Benign focal epilepsy of childhood (disorder)	On examination - salaam attack (disorder)		
	Somatosensory epilepsy (disorder)	Myoclonic encephalopathy (disorder)		
	Partial epilepsy with autonomic symptoms (disorder)	Cryptogenic generalized epilepsy (disorder)		
	Visual reflex epilepsy (disorder)	Cryptogenic West syndrome (disorder)		

	Simple partial epileptic seizure (disorder)	Symptomatic West syndrome (disorder)		
	Visceral epilepsy (disorder)	Myoclonic astatic epilepsy (disorder)		
	Localization-related symptomatic epilepsy (disorder)	Myoclonic absence epilepsy (disorder)		
	Localization-related symptomatic epilepsy with specific precipitant (disorder)	Idiopathic myoclonic epilepsy (disorder)		
	Motor cortex epilepsy (disorder)	Symptomatic myoclonic epilepsy (disorder)		
	Motor epilepsy (disorder)	Eyelid myoclonus with absences (disorder)		
	Jacksonian, focal or motor epilepsy (disorder)	West syndrome (disorder)		
	Visual epilepsy (disorder)	Perioral myoclonia with absences (disorder)		
	Epilepsy of infancy with migrating focal seizures (disorder)	Infantile spasm and broad thumb syndrome (disorder)		
	Familial focal epilepsy with variable foci (disorder)	Atonic epilepsy (disorder)		
	Recurrent benign focal seizures of childhood (disorder)	Sodium voltage-gated channel alpha subunit 8-related epilepsy with encephalopathy (disorder)		
	Benign infantile focal epilepsy with midline spikes and waves during sleep (disorder)	Petit-mal epilepsy (disorder)		
	Benign partial epilepsy with secondarily generalized seizures in infancy (disorder)			
	Autosomal dominant epilepsy with auditory features (disorder)			
	Extratemporal epilepsy (disorder)			
	Partial epilepsy with impairment of consciousness (disorder)			
	Temporal lobe epilepsy (disorder)			
	Psychosensory epilepsy (disorder)			
	Mesiobasal limbic epilepsy (disorder)			

	Lateral temporal epilepsy (disorder)			
	Psychomotor epilepsy (disorder)			
	Epilepsy characterized by intractable complex partial seizures (disorder)			
	Benign familial mesial temporal lobe epilepsy (disorder)			
	Benign partial epilepsy of infancy with complex partial seizures (disorder)			
	Mesial temporal lobe epilepsy with hippocampal sclerosis (disorder)			
	Infantile-onset mesial temporal lobe epilepsy with severe cognitive regression (disorder)			
	Familial temporal lobe epilepsy (disorder)			
	Familial mesial temporal lobe epilepsy with febrile seizures (disorder)			
	Partial epileptic seizure of parietal lobe with impairment of consciousness (disorder)			
	Partial epileptic seizure of frontal lobe with impairment of consciousness (disorder)			
	Partial epileptic seizure of occipital lobe with impairment of consciousness (disorder)			
	Partial epileptic seizure of temporal lobe with impairment of consciousness (disorder)			
	Intractable partial temporal lobe epilepsy with impairment of consciousness (disorder)			
	Intractable partial parietal lobe epilepsy with impairment of consciousness (disorder)			
	Intractable partial frontal lobe epilepsy with impairment of consciousness (disorder)			
	Intractable partial occipital lobe epilepsy with impairment of consciousness (disorder)			
	Decision-making epilepsy (disorder)			
	Epilepsia partialis continua (disorder)			

	Refractory epilepsia partialis continua (disorder)			
	Partial frontal lobe epilepsy (disorder)			
	Cursive (running) epilepsy (disorder)			
	Partial occipital lobe epilepsy (disorder)			
	Partial parietal lobe epilepsy (disorder)			
	Localization-related epilepsy (disorder)			
	Supplementary motor epilepsy (disorder)			
	Secondary reading epilepsy (disorder)			
	Writing epilepsy (disorder)			
	Eating epilepsy (disorder)			
	Toothbrushing epilepsy (disorder)			
	Aquagenic epilepsy (disorder)			
	Refractory localization-related epilepsy (disorder)			
	Refractory parietal lobe epilepsy (disorder)			
	Autosomal dominant nocturnal frontal lobe epilepsy (disorder)			
	Hot water reflex epilepsy (disorder)			
	Thinking epilepsy (disorder)			
	Startle epilepsy (disorder)			
	Micturition induced epilepsy (disorder)			
	Orgasm induced epilepsy (disorder)			
	Audiogenic epilepsy (disorder)			
	Celiac disease with epilepsy and cerebral calcification syndrome (disorder)			
	Infant epilepsy with migrant focal crisis (disorder)			
	Infantile convulsion and choreoathetosis syndrome (disorder)			

~ The End ~