Ascertaining Pain in Mental Health Records

Combining Empirical and Knowledge-Based Methods for Clinical Modelling of Electronic Health Record Text

Chaturvedi, Jaya

Awarding institution:
King's College London

Ascertaining Pain in Mental Health Records: Combining Empirical and Knowledge-Based Methods for Clinical Modelling of Electronic Health Record Text

Jaya Chaturvedi

Supervisors:
Dr Angus Roberts
Prof Robert Stewart
Dr Sumithra Velupillai

Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London

A thesis submitted in fulfilment of the requirements for the degree of *Doctor of Philosophy* in Health Informatics

September 2023

Declaration

I hereby declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contributions and those of the other authors to this work have been explicitly indicated in the relevant chapters. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Jaya Chaturvedi

15/09/2023

Acknowledgements

I want to acknowledge and thank the Department of Biostatistics and Health Informatics at King's College London for funding my PhD research and supporting me at every step of the way.

My supervision team - Dr Angus Roberts, Prof Robert Stewart, and Dr Sumithra Velupillai - have helped me self-reflect at various stages of the PhD, making me better understand the research process and become a better researcher. They have presented me with numerous opportunities through which I have grown and expanded my comfort zone. I would also like to thank Dr Brendon Stubbs for his invaluable input into parts of my project, along with Dr Eugenia Romano and Dr Ruimin Ma for introducing me to the patient and public involvement group, which has enhanced my project. Members of the CRIS team – Daisy Kornblum, Debbie Cummings, Megan Pritchard, Amelia Jewell, Anna Kolliakou, Jyoti Verma, Matthew Broadbent, and David Chandran - have been very helpful with the data access and extraction process. The medical student annotators - Natalia Chance, Veshalee Vernugopan, and Luwaiza Mirza - have been excellent at completing more annotations than I had anticipated, which has been essential in building models that perform well. Alberto Bernardi, a member of the AmpliGraph team, has helped me through the errors and barriers I faced when using the AmpliGraph Python library.

The Drive-Health cohort has provided support and a safe space to share our struggles and successes. My peers Aurelie Mascio, Naoko Skada, Tom Searle, Zeljko Kraljevic, Anthony Shek, Diana Shamsutdinova, Lifang Li, Tao Wang and Yamiko Msosa have always been around to help me with the technical issues I have faced and been an excellent sounding board through these three years. I also want to thank Fernando Lanza and Ruth Fernandes

for being patient with all my queries and helping me on multiple occasions so that I could go out there and present at different conferences without the stress of travel and stay.

Lastly, I would like to thank my close friends, my family (especially my nephew Neel and niece Miraya for bringing me joy at times of stress), my boyfriend Luke (and his family), for their words of encouragement, especially at moments of self-doubt when I needed them the most. I could not have done this without their support.

Abstract

In recent years, state-of-the-art clinical Natural Language Processing (NLP), as in other domains, has been dominated by neural networks and other statistical models. In contrast to the unstructured nature of Electronic Health Record (EHR) text, biomedical knowledge is increasingly available in structured and codified forms, underpinned by curated databases, machine-readable clinical guidelines, and logically defined terminologies. This thesis examines the incorporation of external medical knowledge into clinical NLP and tests these methods on a use case of ascertaining physical pain in clinical notes of mental health records.

Pain is a common reason for accessing healthcare resources and has been a growing area of research, especially its impact on mental health. Pain also presents a unique NLP problem due to its ambiguous nature and the varying circumstances in which it can be used. For these reasons, pain has been chosen as a use case, making it a good case study for the application of the methods explored in this thesis. Models are built by assimilating both structured medical knowledge and clinical NLP and leveraging the inherent relations that exist within medical ontologies. The data source used in this project is a mental health EHR database called CRIS, which contains de-identified patient records from the South London and Maudsley NHS Foundation Trust, one of the largest mental health providers in Western Europe.

A lexicon of pain terms was developed to identify documents within CRIS mentioning pain-related terms. Gold standard annotations were created by conducting manual annotations on these documents. These gold standard annotations were used to build models for a binary classification task, with the objective of classifying sentences from the clinical text as “relevant”, which indicates the sentence contains relevant mentions of pain, i.e., physical pain affecting the patient, or “not relevant”, which indicates the sentence does not contain mentions

of physical pain, or the mention does not relate to the patient (ex: someone else in physical pain). Two models incorporating structured medical knowledge were built:

1. a transformer-based model, SapBERT, that utilises a knowledge graph of the UMLS ontology, and

2. a knowledge graph embedding model that utilises embeddings from SNOMED CT, which was then used to build a random forest classifier. This was achieved by modelling the clinical pain terms and their relations from SNOMED CT into knowledge graph embeddings, thus combining the data-driven view of clinical language, with the logical view of medical knowledge.

These models have been compared with NLP models (binary classifiers) that do not incorporate such structured medical knowledge:

1. a transformer-based model, BERT_base, and

2. a random forest classifier model.

Amongst the two transformer-based models, SapBERT performed better at the classification task (F1-score: 0.98), and amongst the random forest models, the one incorporating knowledge graph embeddings performed better (F1-score: 0.94). The SapBERT model was run on sentences from a cohort of patients within CRIS, with the objective of conducting a prevalence study to understand the distribution of pain based on sociodemographic and diagnostic factors.

The contribution of this research is both methodological and practical, showing the difference between a conventional NLP approach of binary classification and one that incorporates external knowledge, and further utilising the models obtained from both these approaches in

a prevalence study which was designed based on inputs from clinicians and a patient and public involvement group. The results emphasise the significance of going beyond the conventional approach to NLP when addressing complex issues such as pain.

Table of Contents

List of Figures and Tables

List of Figures

List of Tables

List of abbreviations

AI: Artificial Intelligence

ANN: Artificial Neural Network

AUC: Area Under the Curve

BERT: Bidirectional Encoder Representations from Transformers

BoW: Bag-of-Words

BRC: Biomedical Research Centre

CAMHS: Child and Adolescent Mental Health Services

CI: Confidence Interval

CNN: Convolutional Neural Network

CRIS: Clinical Record Interactive Search

CUI: Concept Unique Identifier

DT: Decision Trees

ECG: Electrocardiogram

EHR: Electronic Health Record

ePAT: Electronic Patient Authored Text

ePJS: Electronic Patient Journey System

GP: General Practitioner

GPT: Generative Pre-training Transformer

GPU: Graphics Processing Unit

HIPAA: Health Insurance Portability and Accountability Act

HoNOS: Health of the Nation Outcomes Scales

HTML: Hypertext Markup Language

HTN: Hypertension

IAA: Inter-Annotator Agreement

IASP: International Association for the Study of Pain

ICC: Intra-Class Correlation

ICD-10: International Classification of Diseases, version 10

KG: Knowledge Graph

KGE: Knowledge Graph Embedding

KNN: K-Nearest Neighbours

LBP: Lower Back Pain

LDN: Lambeth DataNet

LL: log-likelihood

LR: Logistic Regression

LSTM: Long Short-Term Memory

LSVC: Linear Support Vector Classifier

MedCAT: Medical Concept Annotation Tool

MeSH: Medical Subject Headings

MI: Mutual Information

MIMIC: Medical Information Mart for Intensive Care

ML: Machine Learning

NB: Naïve Bayes

NER: Named Entity Recognition

NHS: National Health Service

NIMH: National Institute of Mental Health

NIHR: National Institute of Health Research

NLP: Natural Language Processing

NLTK: Natural Language Toolkit

NN: Neural Network

NOS: Not Otherwise Specified

NSAID: Non-steroidal anti-inflammatory drugs

OR: Odds Ratio

PAS: Patient Administration System

PMC: PubMed Central

PPI: Patient and Public Involvement

RF: Random Forest

ROC: Receiver Operating Curve

SapBERT: Self-alignment Pre-training for BERT

SCTID: SNOMED CT Identification

SGD: Stochastic Gradient Descent

SLaM: South London and Maudsley NHS Foundation Trust

SMI: Severe Mental Illness

SMOTE: Synthetic Minority Oversampling TEchnique

SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

SQL: Structured Query language

SVC: Support Vector Classification

SVM: Support Vector Machine

TFIDF: Term Frequency–Inverse Document Frequency

UMLS: Unified Medical Language System

VA: Veteran Affairs

WHO: World Health Organisation

Disseminations

The publications and presentations that reported the research described in this thesis and related research in which I have collaborated are listed here, organised by journal articles, conference papers, posters, and presentations.

Journal articles

As first author:

Chaturvedi, J., Mascio, A., Velupillai, S. U., & Roberts, A. (2021). **Development of a Lexicon for Pain**. Frontiers in Digital Health, 193.

Link: https://www.frontiersin.org/articles/10.3389/fdgth.2021.778305/full

Chaturvedi J, Chance, N., Mirza, L., Vernugopan, V., Velupillai, S. U., Stewart, R., & Roberts, A. (2023). **Development of a Corpus Annotated with Mentions of Pain in Mental Health Records: A Natural Language Processing Approach.** JMIR Formativ Res Published Online First: 19 January 2023. doi:10.2196/45849

Link: https://formative.jmir.org/2023/1/e45849

Chaturvedi, J., Shamsutdinova, D., Zimmer, F., Velupillai, S., Stahl, D., Stewart, R., & Roberts, A. (2023). **Sample Size in Natural Language Processing within Healthcare Research.** Preprint.

Link: https://doi.org/10.2139/ssrn.4553964

Chaturvedi, J., Ashworth, M., Stewart, R., & Roberts, A. (2023). **Distributions of Recorded Pain in Mental Health Records: A Natural Language Processing Based Study**. Accepted. BMJ Open.

As contributor:

Mirza, L., Das-Munshi, J., Chaturvedi, J., Wu, H., Kraljevic, Z., Searle, T., Shaari, S., Mascio, A., Skiada, N., Roberts, A., Bean, D., Stewart R., Dobson R. & Bendayan R. (2021). **Investigating the Association between Physical Health Comorbidities and Disability in Individuals with Severe Mental Illness**. European Psychiatry, 1-33.
Link: https://pubmed.ncbi.nlm.nih.gov/34842128/

Bendayan, R., Kraljevic, Z., Shaari, S., Das-Munshi, J., Leipold, L., Chaturvedi, J., Mirza, L., Aldelemi, S., Searle, T., Chance, N., Mascio, A., Skiada, N., Wang, T., Roberts, A., Stewart, R., Bean, D., & Dobson, R. (2022). **Mapping multimorbidity in individuals with schizophrenia and bipolar disorders: evidence from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register.** BMJ Open, 12(1), e054414. https://doi.org/10.1136/bmjopen-2021-054414
Link: https://bmjopen.bmj.com/content/12/1/e054414.full

Francis, E. R., Fonseca de Freitas, D., Colling, C., Pritchard, M., Kadra-Scalzo, G., Viani, N., Chaturvedi, J., Denee, T. R., Kerr, C., Desai, M., Scott, G., Shetty, H., Broadbent, M., Chandran, D., Downs, J., Velupillai, S., Khondoker, M., Stewart, R., Dutta, R., & Hayes, R. D. (2021). **Incidence of suicidality in people with depression over a 10-year period treated by a large UK mental health service provider.** BJPsych Open, 7(6). https://doi.org/10.1192/bjo.2021.1054

Link: https://www.cambridge.org/core/journals/bjpsych-open/article/incidence-of-suicidality-in-people-with-depression-over-a-10year-period-treated-by-a-large-uk-mental-health-service-provider/68ADF89E00E729D8F2CEE608DFB9D021

Bendayan, R., Wu, H., Chaturvedi, J., Kraljevic, Z., Mascio, A., Searle, T., Dobson, R. (2022). **Using Natural Language Processing to Identify Multimorbidities in Individuals with Severe Mental Illness.** Journal of Biomedical Informatics. *Submitted.*

Ariño H, Bae SK, Chaturvedi J, Wang T and Roberts A (2023) **Identifying encephalopathy in patients admitted to an intensive care unit: Going beyond structured information using natural language processing**. Front. Digit. Health 5:1085602. doi: 10.3389/fdgth.2023.1085602.

Patel, R., Brinn, A., Irving, J., Chaturvedi, J., Gudiseva, S., Correll, C. U., … McGuire, P. (2023). **Oral and long-acting injectable antipsychotic discontinuation and relationship to side effects in people with first episode psychosis: a longitudinal analysis of electronic health record data.** Therapeutic Advances in Psychopharmacology, 13, 20451253231211576. https://doi.org/10.1177/20451253231211575

Conference papers

As first author:

Chaturvedi, J., Velupillai, S., Stewart, R., and Roberts, A. (2024). **Identifying mentions of**

pain in mental health records text: A natural language processing approach. Studies in Health Technology and Informatics, 310, 695–699. https://doi.org/10.3233/SHTI231054

Chaturvedi J, Wang T, Velupillai S, Stewart R, Roberts A. **Development of a Knowledge Graph Embeddings Model for Pain**. AMIA Annu Symp Proc. 2024 Jan 11;2023:299-308. PMID: 38222382; PMCID: PMC10785867.
Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10785867/

As contributor:

Bendayan, R., Wu, H., Kraljevic, Z., Stewart, R., Searle, T., Chaturvedi, J., Das-Munshi, J., Ibrahim, Z., Mascio, A., Roberts, A., Bean, D., & Dobson, R. (2020). **Identifying physical health comorbidities in a cohort of individuals with severe mental illness: An application of SemEHR**. ArXiv. https://doi.org/10.48550/arxiv.2002.08901. HealTAC 2020.
Link: https://arxiv.org/abs/2002.08901

Conference posters

As first author:

Chaturvedi, J., Shamsutdinova, D., Velupillai, S. U., Stahl, D., Stewart, R., & Roberts, A. (2022). **Sample Size in Healthcare NLP.** HealTAC 2022
Link: https://healtac2022.github.io/programmes/

Chaturvedi, J., Chance, N., Mirza, L., Vernugopan, V., Velupillai, S. U.,Stewart, R., & Roberts, A. (2022). **Annotation of Mentions of Pain in Mental Health Records.** IASP 2022 World

Congress on Pain.

Link: https://iaspworldcongress2022.org/schedule-at-a-glance/

Chaturvedi, J., Shamsutdinova, D., Zimmer F., Velupillai, S. U., Stahl, D., Stewart, R., & Roberts, A. (2023). **Sample Size in Healthcare NLP.** HealTAC 2023

Link: http://healtex.org/healtac-2023/programme/

Chaturvedi, J., Velupillai, S. U., Stewart, R., & Roberts, A. (2022). **Recording of Pain within Mental Health Records Text.** APDP 2023 Conference.

Link: https://s3.eu-west-2.amazonaws.com/sixcircles.production/uploads/site_id_87/2023/05/23/document/5sKad0HsBS0V3Xf2WAAqOQe1iLJnbQTUoMzIoX7TovVnRcRmT6KaL6WdCMXuyVkm.pdf

Chaturvedi, J., Velupillai, S. U., Stewart, R., & Roberts, A. (2023). **Distribution of Pain within a Mental Health Records Database.** EFIC European Pain Conference 2023.

Conference presentations

As first author:

Chaturvedi, J., Mascio, A., Velupillai, S. U., & Roberts, A. (2021). **Development of a Lexicon for Pain.** HealTAC 2021

link: http://healtex.org/healtac-2021/programme/

Chaturvedi, J., Velupillai, S. U., Stewart, R., & Roberts, A. (2022). **Combining Empirical and Knowledge-based Methods for Clinical Modelling of Electronic Health Record Text.** Doctoral Consortium at the 20th International Conference on Artificial Intelligence in Medicine 2022.

Link: https://aime22.aimedicine.info/index.php/program/doctoral-consortium-program

Chaturvedi, J., Velupillai, S. U., Stewart, R., & Roberts, A. (2022). **Extracting Information About Pain from Mental Health Records.** EPA Section of Epidemiology & Social Psychiatry 20th Biennial Congress 2022, Cambridge

link: https://www.psychepi.org/congress/wp-content/uploads/sites/2/2022/08/EPA-Section-20th-Congress-Full-Programme-v1.3.pdf

CHAPTER 1: Introduction

1.1 Research problem

1.1.1 Background and Problem Definition

Electronic Health Record (EHR) databases are a longitudinal record of patients' health and healthcare (Horton et al., 2019). They can be defined as electronic versions of patient medical histories maintained by healthcare providers (Keshta & Odeh, 2020), where patient data are collected during routine healthcare delivery (Denaxas & Morley, 2015). A shift from paper-based patient records to electronic records was motivated by the increasingly cumbersome process of maintaining extensive paper trails in former systems (Keshta & Odeh, 2020). EHR databases have been a much more effective way of recording patient histories, and access to longitudinal records for patients has improved healthcare quality (Carey et al., 2016). These EHR databases, in an anonymised format, are often used for research purposes, although not all systems permit this type of use. One reason for their use in research is that they give access to large amounts of patient data which might otherwise not be feasible to obtain, and which can be used for population-based studies. Another reason is the belief that EHRs contain more accurate and detailed information on the true clinical states of patients compared to administrative databases, such as PAS (Patient Administration System), which are solely used for billing purposes. The latter do not contain information such as symptoms, differential diagnoses, laboratory results, or other patient behaviours such as smoking and drug use (Horton et al., 2019). However, since research is not the primary purpose of EHRs, they are heterogeneous, incomplete and noisy (Kim et al., 2019). While they contain large amounts of data, this is frequently in varying and incompatible formats, such as information

structured and coded using some form of national or local coding system or ontology, combined with unstructured free text such as clinical notes, pathology and radiology reports. The potential of using EHR “big data” has been recognised in numerous research projects over the past decades (Carey et al., 2016; Denaxas & Morley, 2015; Jensen et al., 2012; Khoury et al., 2013; Kim et al., 2019; Weber et al., 2014), and they are increasingly used for phenotyping, identification of cohorts for trial eligibility, and other similar studies (Shivade et al., 2014).

In addition to the challenges of heterogeneous and messy EHR data, researchers face ethical and legal issues that might limit access to such data (Kush et al., 2008; Taylor, 2008). In a UK context, this includes considering the national data opt-out applied in England, which has been in action within the NHS since 2018 and made compulsory in 2022. This allows patients to opt out of their data being used for research purposes (NHS Digital, 2022a), ensuring individual patients' right to privacy. Data governance also leads to specific requirements for the security and safekeeping of EHR data. These requirements are threefold - physical security, whereby unauthorised persons are not allowed access to the data; technical security, where firewalls and encryptions are used to protect the data from technical breaches and malware/virus attacks; and administrative security, which includes regular audits of the data and ensuring contingency plans are in place (Keshta & Odeh, 2020).

A large proportion of clinical information is stored within the text fields of EHRs (Kharrazi et al., 2018). These text fields include referral and correspondence letters to other health services, as well as clinical notes from face-to-face interactions with patients. The writer assumes a certain amount of background knowledge in the reader, without which they may be unable to understand the implications of what is being said. For example, in the text, “patient presented with productive cough, left flank rash and epigastric pain. She has been

diagnosed with shingles and left lobe pneumonic process, and is on therapy for both", the reader needs domain expertise to understand that the productive cough is potentially due to pneumonia affecting the left lobe of her lungs, and the rash on her side could be due to the shingles. She is receiving treatment for both of these conditions. Such background knowledge is inherently lacking in machines. There is, therefore, a need to incorporate world knowledge to assist machines in better understanding the meaning beyond the words in the document. State-of-the-art natural language processing (NLP) approaches are frequently used to extract a variety of information from this free text, such as symptoms, medications, and diagnoses (Colling et al., 2020; Viani et al., 2019, 2020, 2021). These NLP approaches follow an empirical data-driven approach, and do not generally incorporate conceptual knowledge (Nadkarni et al., 2011). Sources of conceptual knowledge are, however, readily available in the clinical domain, including structured knowledge such as clinical guidelines, biomedical knowledge such as DrugBank (Wishart et al., 2018), clinical vocabularies such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) (Stearns et al., 2001) and collections of terminological resources such as UMLS (Unified Medical Language System) (Bodenreider, 2004). NLP approaches use statistical methods, while structured knowledge, in contrast, is based on logic-driven structures and representations (Horrocks, 2005). The language models used in current state-of-the-art NLP (such as BERT (Bi-directional Encoder Representations from Transformers)) (Devlin et al., 2018) and GPT-3 (Generative Pre-trained Transformer, an autoregressive language model) (Brown et al., 2020)) are trained on the linguistic forms of words but do not truly understand the meanings inherent in the words, as described in the Octopus test by (Bender & Koller, 2020). For this reason, I hypothesise that incorporating structured medical knowledge into NLP of the EHR text will improve the performance of clinical NLP tasks. To test this hypothesis, this thesis examines the interplay between NLP and structured knowledge, using the extraction of information about pain from

mental health records as an example, and testing the impact this has on downstream NLP and health research.

Pain has been chosen as a use case because it is a major global healthcare problem, with chronic pain affecting one in five adults (International Association for the Study of Pain, 2005). The International Association for the Study of Pain (IASP) defines pain as "*an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage*." (International Association for the Study of Pain, 2020). Pain is known to have a strong relation with negative emotions, which can lead to damaging consequences (Heintzelman et al., 2013). This is worsened for people suffering from persistent pain. It can also lead to long-term mental health effects such as the 'secondary pain effect', which encapsulates strong feelings towards the long-term implications of suffering from pain (Heintzelman et al., 2013). Pain has been an active area of research, especially since the onset of the crisis of opioid use in the United States (Howard et al., 2018). Pain also has a significant impact on the healthcare system and society in terms of costs related to medical care (Groenewald et al., 2014) and loss of productivity (Rayner et al., 2016). A committee reviewing the public health significance of pain in the United States found that the total cost to society was greater than those estimated for heart disease, cancer or diabetes (Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education, 2011).

Pain is a subjective multidimensional experience and acts as an early warning system for potential tissue damage, triggering a response to avoid further harm (Tracey, 2016). The neurophysiology of pain involves transmission along dedicated neural pathways from peripheral nociceptors (specialised nerve endings) to the spinal cord and brain. Nociceptive fibres detect thermal, mechanical, or chemical stimuli and convey signals to the central

nervous system (Aguggia, 2003). Pain is a cardinal symptom of inflammatory disorders (Punchard et al., 2004) and is clinically highly prevalent across many acute and chronic conditions. Pain is commonly experienced by patients with cancer (Snijders et al., 2023), with somatic pain (described as aching, gnawing, throbbing, or cramping) being the most common type, caused due to bone metastases, i.e. cancer cells spreading from their original site to a bone. Nociceptive pain (caused by damage to tissues) is common in inflammatory disorders like arthritis, and neuropathic or nerve pain, resulting from dysfunction of the nervous system itself, manifests in disorders like diabetes, nerve injury, and stroke. In addition, headaches and migraines also have a high prevalence and are linked to neurovascular origins (Carver & Foley, 2003). Pain, with its diverse presentations within healthcare, significantly contributes to disability and reduced quality of life.

Pain management requires a multimodal approach, including pharmacological agents like non-steroidal anti-inflammatory drugs (NSAIDs) and opioids, physical therapy, psychological techniques, and interventional procedures (Cuomo et al., 2019). However, many cases of pain remain untreated, necessitating a better understanding of underlying mechanisms and predictive biomarkers. Further research is needed to optimise therapeutic strategies and account for individual variability in the pain experience.

Several researchers have looked at the link between pain and mental health. For example, a conceptual framework developed by Merlin et al. (2014) illustrated the multidimensional nature of pain and its relationships to social, psychological and biological factors (Merlin et al., 2014). Additionally, a high co-occurrence of pain and mental health disorders has been established and is known to be linked to increased disability and impairment (Vinall et al., 2016). Several other studies have also found associations between pain and mental health issues such as depression (Blumer & Heilbronn, 1982; Eisendrath, 1995; Gureje et al., 1998)

and severe mental illnesses (SMI) (Baughman et al., 2016; Birgenheir et al., 2013; Brooks et al., 2018; Stubbs et al., 2014). Additionally, the World Health Organisation (WHO) conducted a 14-nation study that found an overlap between pain and mental health issues (Gureje et al., 1998).

Patients with SMI are known to have poorer physical health and a higher mortality rate compared to patients without SMI, predominantly due to poor physical health (Onwumere et al., 2022). Furthermore, pain may be under-recognised in the SMI population due to the presentation of pain in this population being more complex than that of the general population (Abplanalp et al., 2020). While previous research has found that patients with schizophrenia experience reduced pain sensitivity compared to the general population (Bleuler, 1988; Potvin & Marchand, 2008), this contradicts other research that found higher severity of bodily pain in patients with schizophrenia when compared to the general population (Strassnig et al., 2003). Baughman et al. (2016), when considering the relationship between comorbidities and disease burden in patients with SMI, found that chronic pain had the greatest disease burden in this population, with chronic pain being two times more common in an SMI sample compared to a non-SMI national comorbidity sample (Baughman et al., 2016). Additionally, a systematic review by Onwumere et al. (2022) found that due to the nature of SMIs, pain communication and assessment are hampered (Onwumere et al., 2022)(Onwumere et al., 2022). Despite this, pain is not regularly assessed and managed in this population (Brendon Stubbs et al., 2015).

Depression is another mental health disorder commonly associated with pain. Eisendrath et al. (1995) conceptualised depression as a contributory cause of pain, a neurobiological companion to pain, as well as a result of inescapable chronic pain (Eisendrath, 1995). A systematic review by IsHak et al. (2018) highlighted the impaired functioning resulting from

the co-occurrence of pain and depression (IsHak et al., 2018), finding that patients with pain and depression experience reduced physical, mental and social functioning when compared to patients with only depression or only pain. In other research, Polatin et al. (1993) found that patients with mental health disorders such as anxiety, depression and personality disorders showed higher prevalence rates for back pain than the general population (Polatin et al., 1993).

As previously mentioned, most patient information is recorded in unstructured clinical narratives within EHR databases (Velupillai et al., 2018), and information on pain is no different. Mental health EHRs are, therefore, a good source of textual information that can help better understand the overlaps between pain and mental health and are, therefore, the particular focus of this thesis. However, these mental health EHRs pose an additional challenge of containing large amounts of pertinent information within the unstructured free-text clinical notes rather than in structured fields. Factors such as clinical uncertainty and the social context of patient care, frequently encountered in mental health care, do not translate well into structured codes and tables. Therefore, free text is often used to record nuanced facts. The information in the clinical notes also provides insight into patient symptoms and care, making it valuable in terms of research (Stewart & Davis, 2016). Mental health notes also importantly contain a lot of sensitive and potentially stigmatising information about the patients' mental state and well-being, thereby intensifying privacy concerns when considering the use of mental health EHRs (Kariotis et al., 2022). There are also concerns about the impact of sharing mental health EHR data on the therapeutic relationships between patients and their clinicians (Kariotis et al., 2022). These concerns highlight the importance of patient involvement in the use of such data and in decisions on whether a particular research project should employ data from such records or not.

However, given the wide range of presentations and experiences of pain, the complex description of pain in these records makes extracting this information a challenging task. Additionally, the complex and ambiguous nature of pain means that in order to better understand what is being implied in the clinical notes, information on several attributes of pain is also required, such as the quality of the pain (sharp, stabbing), pain-related medications, the relationship between the pain and body parts, and so on. This information goes beyond what is explicitly mentioned in the notes and depends on some degree of domain knowledge. There is extensive information about these pain attributes within structured medical knowledge resources, which could complement the data within clinical notes. For example, SNOMED CT (Stearns et al., 2001) is one such resource that consists of concepts in a hierarchical structure and includes structured knowledge on pain and its relationship with other clinical concepts. Therefore, it can be used to provide additional relational information such as meronyms, hypernyms and hyponyms of the pain entities. Modelling SNOMED CT as a knowledge graph could facilitate this approach.

The mental health EHR database used in this project is CRIS (Clinical Records Interactive Search), a de-identified database of EHRs from The South London and Maudsley NHS Foundation Trust (SLaM) (Stewart et al., 2009). SLaM is one of the largest mental healthcare providers in Western Europe, providing services to four London boroughs - Lambeth, Southwark, Lewisham and Croydon (Jackson et al., 2017). The CRIS database was developed in 2008 and is managed through a robust process by a patient-led oversight committee. Ethics approval for CRIS has been granted by the Oxford C Research Ethics Committee (reference 23/SC/0257). Stewart et al. (2009) have outlined the technical architecture and the various safety measures employed in constructing this infrastructure (Stewart et al., 2009). Data from CRIS has been used for numerous epidemiological studies (Biondo et al., 2022; Govind et al., 2022; Ma, Romano, Davis, et al., 2022; Ma, Perera, et al.,

2022; Widnall et al., 2022), many of which have utilised NLP approaches to harness the information within the clinical notes and attachments (Bendayan et al., 2020, 2022; Botelle et al., 2022; Chaturvedi et al., 2019; Mirza et al., 2021; Viani et al., 2020).

While CRIS provides patient data from secondary care, a related data source for patient information from primary care, Lambeth DataNet (LDN) (N H S, 2021a), is also used in this research. As with CRIS, patients can opt out of being included in the LDN by informing their GP (General Practitioner) practices. LDN contains electronic records of over 300,000 patients spanning over 40 GP practices within the London Borough of Lambeth (Dorrington et al., 2021). Records within LDN are pseudonymised by replacing patient NHS (National Health Service) numbers to ensure anonymity. LDN does not provide access to clinical notes. Despite this, LDN has been used in various epidemiological studies (Catalao et al., 2021; Davis et al., 2021; Dorrington et al., 2021; Rowlands et al., 2018; Woodhead et al., 2014). There is an existing patient-level linkage in place between LDN and CRIS.

1.1.2 Aims

A conceptual diagram describing the project is shown in Figure 1.1, with the key aims summarised below:

1. Development of an NLP application to classify sentences as containing mentions of physical pain or not.

2. Linking pain entities within the clinical text to compositional structured knowledge, such as SNOMED CT modelled as knowledge graphs, to make use of the additional relations between the concepts for better utilisation of such information in the classification of sentences within the clinical text.

3. Test the impact of such incorporated knowledge, compared to baseline NLP methods, by using NLP-derived information in a downstream epidemiological study.

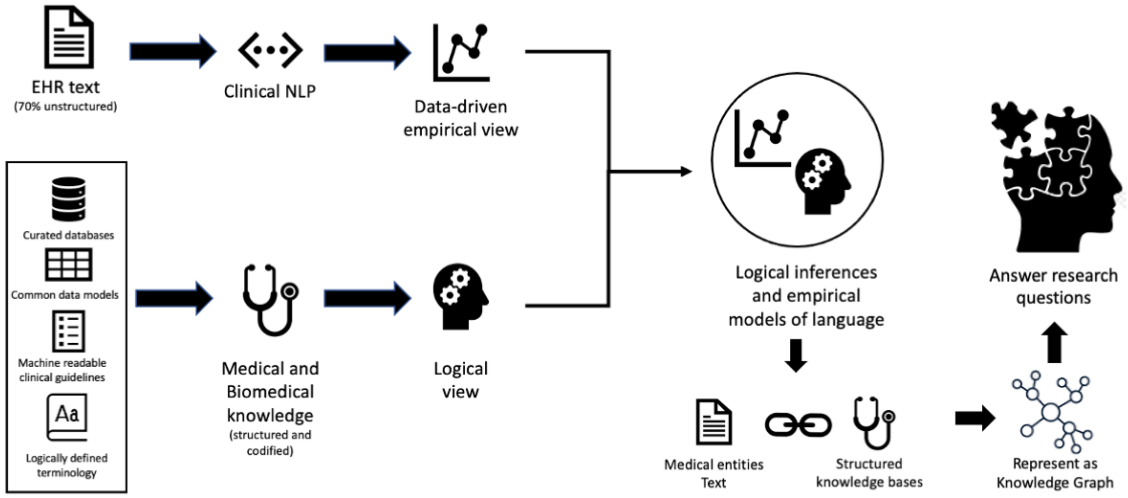


Figure 1.1. Conceptual diagram of the project representing the incorporation of structured knowledge into clinical language to improve extraction of information from free-text clinical notes

The figure shows the combination of EHR text with structured biomedical knowledge, thereby incorporating/linking the logical inferences of the structured knowledge with the empirical models of language from EHR text, which is then represented as knowledge graphs which will aid in answering research questions.

While these methods are being applied to a pain-based use case, they can also be generalised to other clinical concepts.

1.1.3 Originality

Applications have previously been developed to classify sentences for the presence or absence of entities of interest from clinical text. These generally utilise classic supervised machine learning approaches, such as Support Vector Machines (SVM) and K-nearest neighbours (KNN), and more recently, deep learning methods, such as transformer models. This project explores methods of incorporating the logical representations from structured

knowledge into such empirical data-driven models of clinical language to improve the performance of models used to extract information from free text. This approach is novel in its application to a classification task applied to mentions of pain in clinical notes within a mental health EHR database.

This will be achieved by linking mentions of clinical entities within the free text of EHRs to the compositional terms in SNOMED CT, which is extensively adopted in healthcare and clinical research across the globe, including the NHS (Benson, 2010). This will provide additional dimensions to the clinical concepts referred to by the text, by utilising the relations between concepts in the structured knowledge. Linking these entities to structured knowledge modelled as knowledge graphs will help establish the relations between concepts, improving the classification of spans of text containing these entities and, therefore, improving EHR analysis for mental health research. To clarify the terminology used within this thesis, ‘entities’ are specific objects or pieces of information mentioned in the text, while ‘concepts’ represent abstract ideas that categorise groups of objects or things. For instance, in the sentence “the patient was diagnosed with arthritis and was prescribed paracetamol since it was painful in the mornings”, the entity “arthritis” falls within the concept of “disease”, and the entity “paracetamol” belongs to the concept of “medication”. Additionally, a “mention” alludes to an entity. In this example, “arthritis” and “it” are mentions that refer to the arthritis entity.

Experiments based on incorporating structured knowledge with clinical notes will be used to measure the impact on the performance of NLP tasks and the health research based on these NLP tasks.

Codes and guidelines have been made available on GitHub (links available as footnotes within various sections of the thesis, as well as at the end of this chapter). The data used is

stored within the hospital's private network and is available to suitable authorised researchers working on projects approved by the oversight committee of the database.

1.1.4 Research Questions

Three research questions will be answered through this project:

1. Does a system that incorporates domain knowledge into an NLP task perform better than a system without that knowledge?

2. Does this method successfully harness the relations that exist between the pain concepts within a structured knowledge resource and translate them to better classification of sentences within the EHR text?

3. Can this approach be harnessed to extract richer information about pain from mental health EHRs, thereby improving the quality of research outputs?

1.1.5 Outline of work carried out

A set of experiments were conducted to understand whether the incorporation of world knowledge in non-augmented, statistical NLP methods led to an improvement in information extraction from the EHR, thereby leading to richer data available for health research conducted based on this extraction, as shown in the experimental framework (Figure 1.2) and example (Figure 1.3 a and b). The approach was applied to the use case of pain in mental health, and a pain-related prevalence study designed in response to the requirements of clinical colleagues and a patient and public involvement (PPI) group using the outputs of this method.

Figure 1.2. Experimental framework

This figure shows an experimental framework which demonstrates that two approaches will be applied to the EHR text – standard NLP approaches and NLP augmented with knowledge – with the aim of extracting pain information from the EHR text. These approaches will be validated and further used in health research.

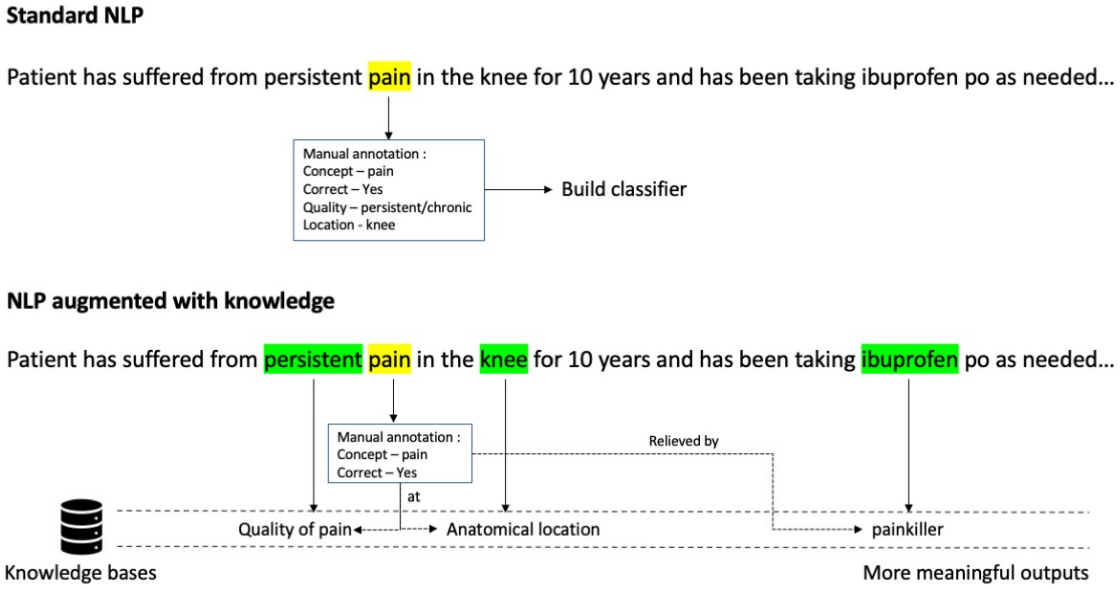


Figure 1.3 (a). Example 1

This figure compares the two approaches mentioned in Figure 1.2, i.e., standard NLP approaches and NLP augmented with knowledge. The Standard NLP example shows the annotated term "pain" (in yellow), and the manual annotations made to record the various attributes associated with the annotated term "pain". The NLP augmented with knowledge example shows other annotations (in green, to highlight that they are additional annotations that were missing in the standard NLP example) that have been identified due to augmentation with structured knowledge bases, leading to more meaningful annotations.

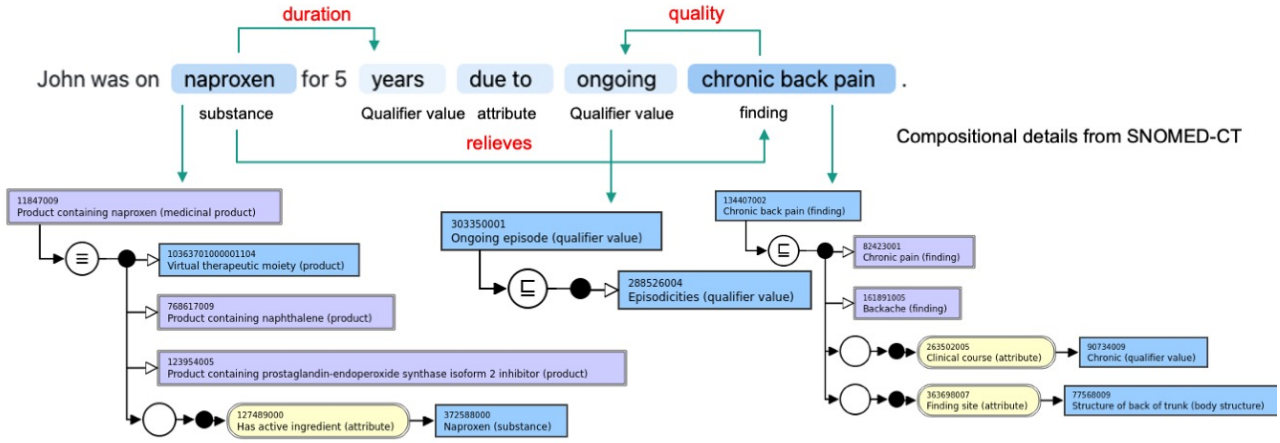


Figure 1.3 (b). Example 2

This figure shows another example of EHR data augmented with structured knowledge (SNOMED CT, in this case). In this example, the annotated words (dark blue indicates words found in SNOMED CT, light blue indicates attributes within the sentence) are linked to SNOMED CT. For demonstration purposes, diagrams for the annotated words were taken from the SNOMED CT browser and are shown here to highlight the compositional data available for each term. Within these diagrams, light blue boxes indicate primitive concepts, purple boxes indicate defined concepts, empty circles mean attribute group elements, dark circles mean conjunction elements, 3 dashes mean equivalent, and the C with an underscore means subsumed by.

Both methods were validated against a dataset that contains independent information about pain, using primary care records within the LDN (N H S, 2021b).

In summary, the following steps were followed to achieve the objectives of this project.

1. A narrative literature review was conducted, detailed in Chapter 2, to explore and understand what has been done in the area thus far. Literature was reviewed around the development and application of NLP on EHRs in research, pain research using EHR data, and the use of NLP methods incorporating external knowledge within healthcare research.

2. To capture the numerous ways in which pain can be described, a lexicon of pain terms was constructed. These terms were obtained from literature, ontologies, and word

embedding models. The terms were reviewed by two clinicians as well as a PPI group. This has been detailed in Chapter 3.

3. This lexicon of pain terms was used to extract documents from a database of mental health EHR records, CRIS (Stewart et al., 2009). The extraction process and the data source have been detailed in Section 5.4.3 of Chapter 5.

4. A simulation study was conducted to determine the appropriate sample size for training data used to train binary classification models. This simulation has been detailed in Chapter 4.

5. Upon extraction of documents from the database, and in order to provide training and evaluation material for supervised models, an annotation exercise was carried out in which three medical student annotators read through each example of a potential pain or related term within the clinical notes and annotated (marked) these terms as relevant, not relevant, or negated. If annotated as relevant, they also added annotations on the pain character, the anatomy affected, and any pain management measures. This process has been detailed in Chapter 5.

6. The gold standard annotations obtained from the annotation process (n=5,644) were used to build two transformer-based classifiers, BERT_base and SapBERT, as well as three other classifiers, SVM, KNN, and Random Forest. The methodology for SVM and KNN has been discussed in Section 6.3 of Chapter 6, with Random Forest described in Section 6.9.1 of Chapter 6. A classifier built specifically to classify whether anatomy is mentioned in relation to pain within the sentences is described in Section 6.9.3 of Chapter 6.

7. In addition to these classifiers, the gold standard annotations were used to construct a knowledge graph embedding model that combined the entities referred to by the pain terms within the annotations with relations for these terms that were obtained from SNOMED CT. The construction of this knowledge graph embedding model is described in Section 7.4 of Chapter 7. This model was then used to build a Random Forest classifier model, which is detailed in Section 7.9 of Chapter 7.

8. The two transformer-based models, BERT_base and SapBERT, were run over a new cohort of patients from the CRIS database to classify sentences within their documents as containing relevant or not relevant mentions of pain. The outputs from both classifiers were compared to each other and to the random forest classifier models built with and without the knowledge graph embedding. The results of these comparisons are detailed in Chapter 8.

9. The CRIS cohort, classified as patients with and without relevant mentions of physical pain by utilising the SapBERT model, was used to conduct a prevalence study. The purpose of this study was to understand the underlying demographic and diagnosis-based distributions between the two classes. In addition to this, the CRIS cohort was compared to a linked subset of patients from LDN, following similar extraction criteria as that of the CRIS cohort, to understand the overlap of mentions of pain between primary and secondary care, as well as form an external validation of the methods employed on the CRIS database. Patient-level linkage already exists between CRIS and LDN. The findings from both these studies can be found in Chapter 9.

10. Finally, Chapter 10 discusses the work, its strengths and limitations, and suggestions for future work.

The aims of this project, the objectives of each chapter with respect to the aims, and the research questions answered by them have been detailed in Table 1.1 below.

Aim	Chapter	Objectives	Research Question Answered
1. Development of an NLP application to identify mentions of pain	3	Develop a lexicon of pain terms	1
	4	Use the lexicon to identify and extract documents from the CRIS database	
	5	Annotation process to generate gold standard annotations from the extracted documents	
	6	Build classifiers using the gold standard annotations	
2. Linking pain entities within clinical text to compositional structured knowledge, SNOMED CT, modelled as knowledge graphs, to make use of the additional relations between concepts	7	Build knowledge graph embedding model by combining the gold standard annotations with relations of the pain terms from SNOMED CT Build a classifier model using this embedding model	1 and 2
3. Conducting experiments to test the impact of this approach when compared to standard NLP approaches by using outputs from these approaches in an	8	Compare the outputs of the various classifier models. Use the best model to run on a new cohort of patients from CRIS	3
	9	Conduct prevalence studies on this	

Aim	Chapter	Objectives	Research Question Answered
epidemiological study.		cohort to better understand the distribution of recorded pain	

Table 1.1. Thesis Chapters and Objectives

This table lists out the aims of the projects, the chapters that cover the aims, the objectives linked to each aim, and the research questions answered by each aim.

1.2 Results and Conclusion

The key contributions from this project are listed below.

1. Development of a comprehensive lexicon for pain, the first such pain lexicon developed. The code and documentation related to the lexicon, as well as the lexicon itself, have been made openly available and can be accessed at https://github.com/jayachaturvedi/pain_lexicon.

2. Development and deployment of an NLP application and framework for pain for the first time in a mental health setting. This framework includes guidelines for the annotation and adjudication process to aid in the development of gold standard annotations, Python code for fine-tuning a pre-trained BERT model and training the non-transformer-based models (SVM, KNN and Random Forest), validation of the model to obtain performance metrics, the classifier models, and code to run the model on unseen documents. All code and documentation have been made available on GitHub at https://github.com/jayachaturvedi/pain_in_mental_health. The pain application is being run on the entire CRIS database for further validation and will be made available as an output for other researchers and future research as part of the NLP deployment services within CRIS.

3. An illustration of the way in which external knowledge contributes to NLP applications. The performance of classification models built with the incorporation of external knowledge was compared to those without, and all models were run on a cohort from CRIS to compare their outputs from the classification task. While this has been tested on a pain use case, the methodology is independent of pain and can easily be replicated for other concepts. Code and documentation for these models have also been made available on GitHub at https://github.com/jayachaturvedi/pain_in_mental_health.

4. A novel comparison of primary and secondary care records in reference to recorded pain experiences was conducted between the CRIS and LDN databases. This comparison provides insights into overlaps or lack thereof between these databases.

5. A website was constructed for this project to generate more engagement with the public and can be accessed at https://sites.google.com/view/pain-mental-health/.

A detailed list of some of the decisions made throughout this project is available in Appendix 1. Since this project involves the training of multiple machine learning models, some of which utilise GPUs (Graphic Processing Units), I would like to acknowledge the carbon emissions from this work that might have contributed to the carbon footprint and, therefore, the environmental impact of such resource-intensive tasks. A calculation of the carbon emissions related to this project, and any carbon offsetting measures are detailed in Appendix 6.

CHAPTER 2: Techniques for Clinical NLP: A Narrative Review

Having set the aims for this project in the previous chapter, this chapter explores the development and application of NLP in EHRs, before reviewing the use of knowledge graph embedding methods within healthcare research, which is of particular relevance to this thesis. The objective of this chapter is to inform on the methods that can be used to achieve the aims of this project.

Keyword searches on PubMed and Semantic Scholar, focusing on EHR and NLP-related keywords [((EHR) OR (Electronic Health Records) OR (Health Records) OR (Medical Records) OR (Electronic Medical Records)) AND ((NLP) OR (Natural Language Processing) OR (Linguistics) OR (Clinical Notes))], retrieved a substantial number (over 20,000) of articles, indicating extensive research activity around applying NLP techniques to EHRs. However, given the vast scale of this literature, this chapter summarises only the studies and developments that seemed most relevant to this thesis in order to provide a narrative overview of the current state of the field. This background and perspective inform the technical approaches taken to address the aims mentioned in Section 1.1.2 of Chapter 1.

Additionally, this chapter explains terminologies, pre-existing methodologies and applications that are used in the subsequent chapters. This chapter also serves as a supplement to some of the upcoming chapters, which contain published papers with literature reviews in their introduction/background sections.

2.1 Natural Language Processing

As discussed in Section 1.1, not all information within EHRs is stored within the structured fields of the databases. A large proportion is stored as unstructured, free-text and is challenging to access. NLP helps overcome these challenges by converting clinical documents into analysable data elements. NLP techniques are generally scalable, adaptable to other similar data, and make the task of data extraction and analysis less time-consuming and less labour-intensive (Kim et al., 2019). This section provides a high-level overview of current NLP techniques.

NLP is a subfield of artificial intelligence (AI) that blends computational linguistics with statistical and machine learning methods to allow for the analysis and processing of text data (I B M, 2021). Clinical NLP (NLP in the healthcare domain) emerged in the 1960s, originating from Zellig Harris's mathematical language theories (Harris, 1968, 1982, 1991). Early NLP systems were developed based on semantic information of near words in sentences (Baud et al., 1992; Haug et al., 1994; Sager et al., 1993), following rules-based approaches using grammatical syntaxes and patterns of regular expressions to match the relevant pieces of text. A review by Wang et al. (2018) spanning from 2009 to 2016 revealed that among the publications surveyed, 65% focused on rule-based information extraction from clinical notes, while 23% employed machine learning techniques (Wang, et al., 2018). Despite the earlier dominance of symbolic, linguistic structure-based and frequency-based approaches (Friedman et al., 2013), recent years have witnessed the gradual adoption of machine learning algorithms, particularly neural networks. Neural networks, based on a simplified model of the human neuron as described by (McCulloch & Pitts, 1943), are a network of simple computing units, each of which takes a vector (array of numbers) off input values, weights each of these inputs, and produces a single output value based on an activation

function applied to the sum of weighted inputs. By connecting multiple such units into layers and other architectures, complex functions of image and language classification can be modelled. The increased adoption of machine learning approaches has been driven by the availability of big data, access to more computational power, and advancements in machine learning methods (Malte & Ratadiya, 2019). These machine learning methods include the development or application of methods that allow computer programmes to derive models from a set of example data, referred to as training data. When such methods are used, the aim is to create models that generalise away from the training data, allowing the programme to make predictions from new unseen data (Nadkarni et al., 2011). This method is categorised as supervised learning and is widely used in NLP for various tasks. These include classification (assigning a label to a unit of text) and sequence labelling, which includes named entity recognition (identifying named entities within the text such as name, location etc.) and part-of-speech tagging (assigning to each word a part of speech such as noun, verb etc.) (Jurafsky & Martin, 2009). Classification can be at the document-level, paragraph-level, sentence-level, or token-level. It can also be binary or multi-class. An essential advantage of sentence-level classification is its focus on immediate context (Yan et al., 2019) (context, in NLP, refers to the words around a term of interest).

State-of-the-art NLP utilises large language models pre-trained on generic text, such as Wikipedia (Devlin et al., 2018), or more domain-specific texts, such as journal articles from PubMed (Gu et al., 2022), clinical notes from a critical care hospital (Huang et al., 2019), or medical ontologies (Liu et al., 2021; Michalopoulos et al., 2021). Pre-training refers to the process of learning representations of meanings of words/sentences, i.e., learning the associations between words that exist within the text, by processing large amounts of text (Jurafsky & Martin, 2009). This is achieved by encoding the relationships and co-occurrences between the words and sentences, with the objective of capturing the syntactic and semantic

relationships. Essentially, the sentences within the text are tokenised (split into individual words or subwords), and each token is assigned a unique numerical vector or “embedding”, which captures the token’s context and meaning in a numerical format, such that tokens that often appear together or have similar roles in sentences will have similar embeddings. The model thus learns the associations and relationships that exist within the text, information which can be used in subsequent NLP tasks. These language models promote the use of shared language representations that capture the semantics of words within generic or domain-specific text. This is possible through a technique called transfer learning, where the knowledge learnt in one task or domain is applied (or transferred) to solve another task. Transfer learning works through a process called fine-tuning, where further training a pre-trained model by adding to the representations already learnt by them, aids in downstream tasks (Jurafsky & Martin, 2023). These approaches are heavily rooted in the aforementioned mathematical theories of language by Zellig Harris, which emphasises breaking down language into mathematical units and analysing how they interact, promoting the capture of the underlying organisation of language.

A popular pre-trained model is BERT (Devlin et al., 2018). BERT utilises a transformer architecture, a type of deep learning architecture that leverages self-attention mechanisms that can capture relationships between words or tokens in a sequence simultaneously, enabling efficient and parallel computation (Vaswani et al., 2017). Self-attention is also known as intra-attention and is a type of attention mechanism that allows the model to focus on, i.e. attend to, different positions of a sequence to compute a representation of this sequence (Vaswani et al., 2017). For example, in the sentence, “The dog jumped over the fence because it was chasing the cat.”, self-attention allows the model to associate “it” with the dog instead of the fence. Transformers are known to handle distant information and are more efficient to implement at scale. They map sequences of input vectors to those of output

vectors of the same length and are made up of stacks of transformer blocks. Each transformer block (Figure 2.1) is a multi-layer neural network that combines simple self-attention linear layers with feedforward networks (information flows unidirectionally from one layer of units to the next) and self-attention layers. They also include residual connections that pass information directly from a lower layer to a higher layer, which improves gradient propagation (i.e., transfer of learning signals through the different layers of the model) and enables more efficient learning. Implementing these residual connections involves adding a layer's input vector to its output vector before passing it forward. In addition to this, the summed input and output vectors are normalised by utilising a "layer normalize" to improve training performance. This improvement is achieved by maintaining the values of a hidden layer in a range that facilitates training (Jurafsky & Martin, 2023).

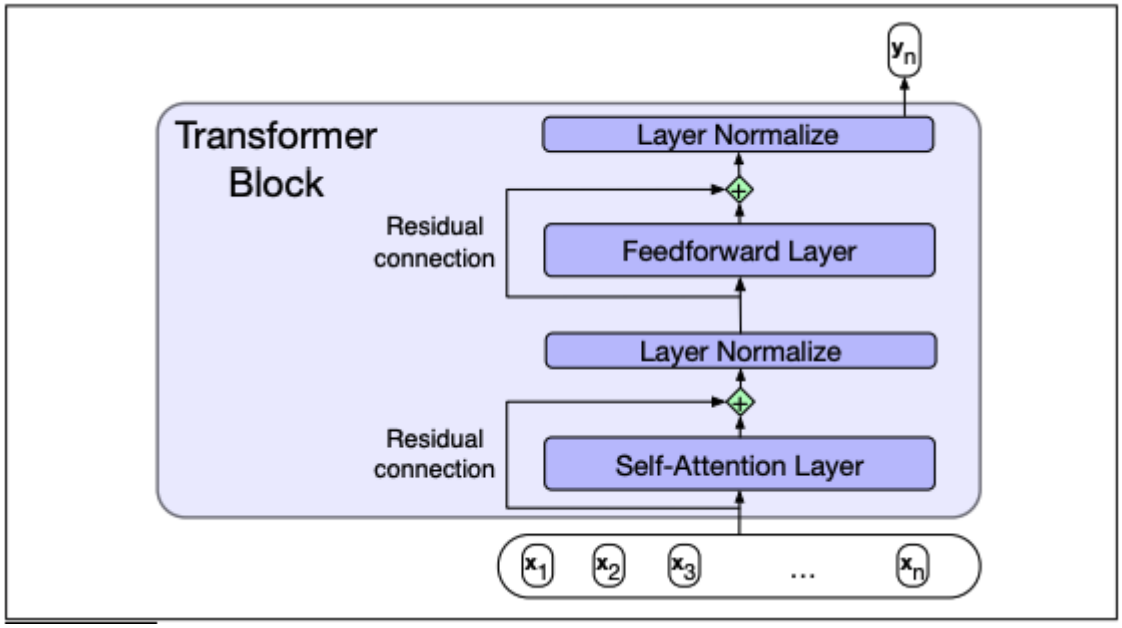


Figure 2.1. Transformer block
(Jurafsky & Martin, 2023)

This figure shows a transformer block within the light purple square, with the dark purple squares indicating the various layers within it. The x components are the inputs entering the transformer block, and the y component is the output exiting the block after passing through all the layers.

The self-attention layer allows a network to extract information from large amounts of text. The flow of information through a single self-attention layer is shown in Figure 2.2, where x_1 to x_5 are the input sequences and are mapped to y_1 to y_5, which are the output sequences.

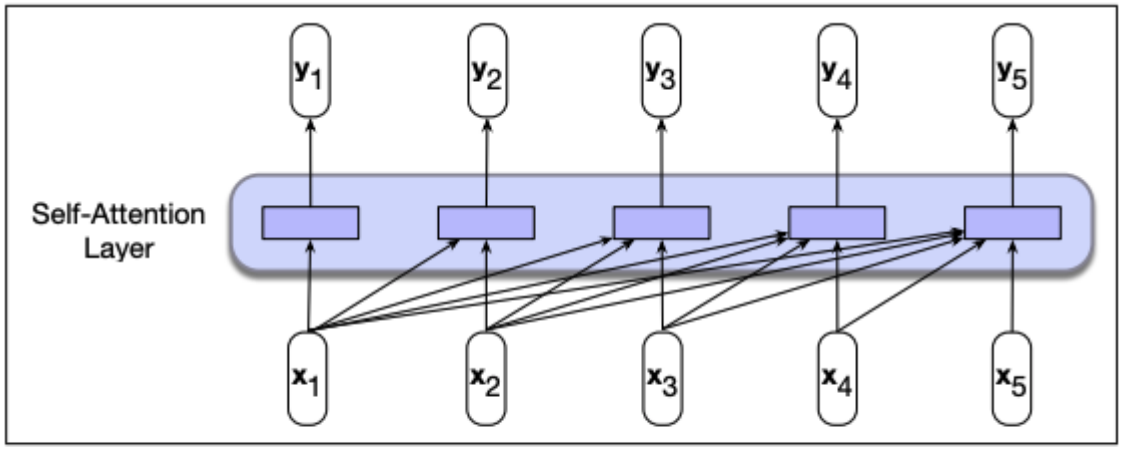


Figure 2.2. The flow of information through a self-attention layer
(Jurafsky & Martin, 2023)

This figure shows the flow of information through a self-attention layer (in purple) within a transformer block. Important to note here is that the information from each input (x_n) flows unidirectionally to generate the outputs (y_n).

When the self-attention layer processes each input item (say x_3), it has access to all the inputs up to and including that input (x_1 to x_3), which allows a comparison of these inputs to reveal their relevance within the current context. These comparisons are then used to compute the output (y_3 in this instance) for the current input (x_3). The simplest form of such comparisons is a dot product. The scores of the dot products are normalised using a SoftMax function to create a vector of weights. A SoftMax function maps values to a probability distribution, with each value within the range of 0 to 1, and all values summing up to 1 (Jurafsky & Martin, 2023).

As mentioned in Section 1.1 of the Introduction chapter, background knowledge is essential to understand what is being said within clinical notes. This is especially true for text about

pain due to its ambiguous nature and the varying scenarios where it can be used. BERT models present a solution to the problem of ambiguity as they learn the context of words based on their surroundings from both directions (not just left to right like other language models). Unlike the example previously discussed (Figure 2.2), BERT allows the self-attention layer to access all the inputs in both directions (x_1 to x_5), so each output vector is contextualised using information from the entire input sequence, as seen in Figure 2.3. Apart from this change, all other features of the BERT model follow the basic transformer architecture.

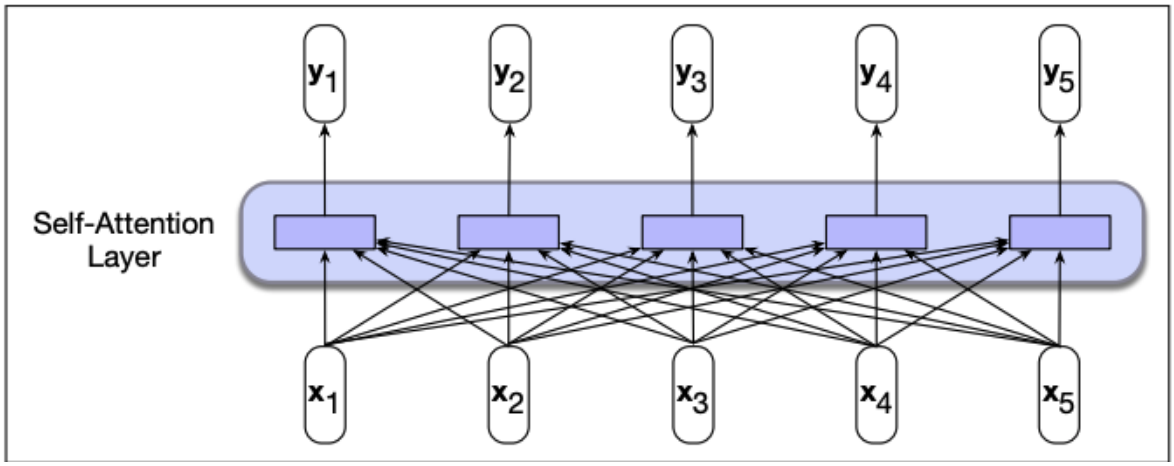


Figure 2.3. Flow of information through a bi-directional self-attention layer
(Jurafsky & Martin, 2023)

Similar to Figure 2.2, this figure shows the flow of information through a self-attention layer (in purple), with the difference being the bi-directional flow of information amongst the inputs (x_n) as seen in BERT.

The transformer architecture is composed of encoders and decoders. Encoders comprise layers of multi-head self-attention mechanisms and fully connected feed-forward networks. These layers process any input sequences and generate representations for them. Importantly, this representation is independent of the output. Decoders are composed of an encoder, plus an additional layer which performs multi-head attention over the output of the

encoder stack. The decoder uses the independent representations generated by the encoder to produce the output sequences. The transformer architecture can, therefore, handle input sequences of a different length than the required output. The BERT representation consists of the parameters learned for an encoder layer. This is because decoders are typically used for sequence generation and language translation tasks, while BERT models are designed for pre-training on unsupervised tasks (such as masked language modelling (MLM), where the BERT model learns to predict masked words in a sentence by examining the context surrounding the word, and next sentence prediction (NSP), where the BERT model predicts if two sentences are consecutive or not) and so do not require the functionalities of a decoder. This simplifies the BERT architecture and reduces the computational load and complexity.

BERT consists of 3 key features:

1. a subword vocabulary of 30,000 tokens - these were generated using the WordPiece algorithm, a large family of subword tokenisation algorithms where if a word does not exist in the vocabulary, the word is divided into subword units by adding a prefix. For example, a word like ‘arthritis’ could be split into ‘art’, ‘##hr’, ‘##itis’

2. a hidden layer of size 768

3. 12 layers of transformer blocks, with each containing 12 multihead attention layers

With a model like BERT, the size of the input layer dictates the complexity of the model and affects the memory and time requirements based on its length. For this reason, a fixed input size of 512 subword tokens is recommended (Devlin et al., 2018). BERT models have been increasingly used in clinical NLP, ranging from tasks such as identification of symptoms (Faris et al., 2022), interpersonal violence (Botelle et al., 2022), adverse drug reactions (Portelli et al., 2021; Wu et al., 2021), and clinical concept extraction (Si et al., 2019).

2.2 Structured Knowledge and Knowledge Graph Embeddings

For many years, integrating domain knowledge into NLP tasks has demonstrated its effectiveness (Azzam et al., 1999; Gaizauskas & Humphreys, 1997; Humphreys et al., 1998), and it is still applicable in current times, reflecting the ongoing recognition of the value that integrating knowledge has on modern NLP techniques. Structured domain knowledge within healthcare is available in various forms - SNOMED CT (controlled clinical vocabulary), UMLS (collection of terminological resources), ICD-10, i.e., International Classification of Diseases version 10 (World Health Organization, 2008) (classification system)), clinical guidelines, common data models, and other biomedical databases. Data within EHR databases are generally linked to some of this structured knowledge, such as SNOMED CT, which is a national requirement within the NHS England electronic patient record systems, for the ease of data sharing and analysis, as well as ease of input by clinicians at the user interface, thereby enabling consistency and accuracy in recording patient data (NHS Digital, 2023). This allows the clinician to type any variation of a condition (such as heart attack, cardiac infarction, MI, or myocardial infarction) that is linked to a single identifier from SNOMED CT. This ensures consistent and accurate data recording, simplifying the exchange of clinical information between systems (NHS Digital, 2023). SNOMED CT follows a compositional, post-coordinated system. Consider, for example, the clinical entity “acute appendicitis”. Within SNOMED CT, this can be represented either as “acute appendicitis” itself or as a combination of “acute inflammation” and the finding site “appendix”, as seen in Figure 2.4.

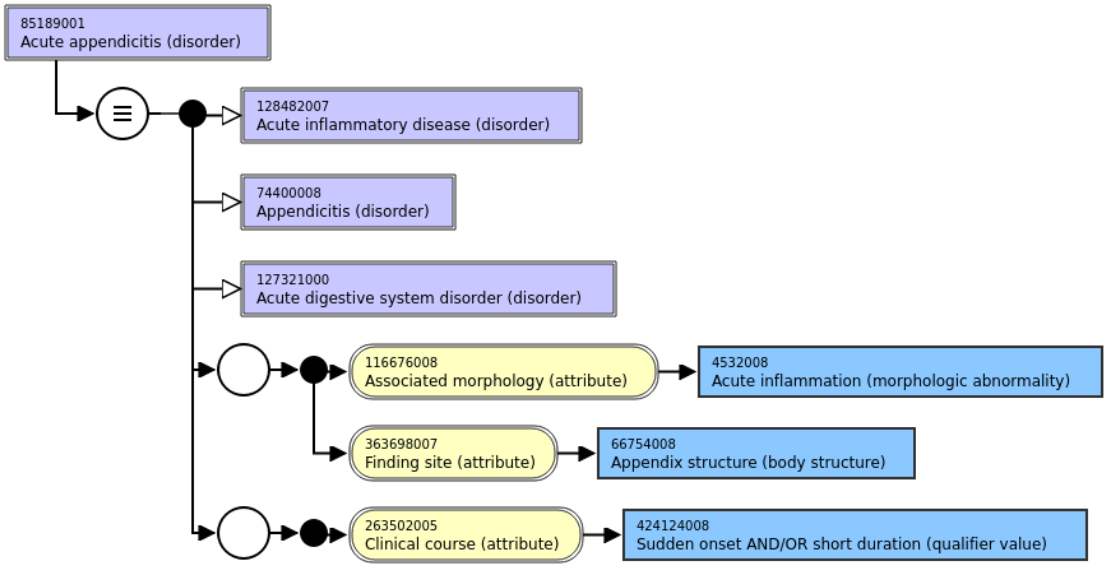


Figure 2.4. Compositional post-coordinated representation of "acute appendicitis" within SNOMED CT

Each number within the boxes indicates the SNOMED CT Identification Number (SCTID) for each concept and sub-concept shown. As seen in Figure 1.3(b), within this diagram, light blue boxes indicate primitive concepts, purple boxes indicate defined concepts, yellow boxes indicate attributes, empty circles mean attribute group elements, dark circles mean conjunction elements, and 3 dashes mean equivalent.

This compositionality is an advantage over other biomedical terminologies in that, in both cases, the result will have the same conceptual and computational meaning (Haendel et al., 2018). Multiple ways of describing the same concept will, therefore, be treated as equivalent in any analysis. Furthermore, similar to the heart attack example mentioned before, SNOMED CT allows for concept normalisation where terms such as "pain", "painful", "dolor", "dolour", and "part hurts" will all link back to the same unique SNOMED CT identification number (SCTID), so regardless of how the pain is mentioned within the text, it is standardised to a particular SCTID, as seen in Figure 2.5. However, this linking based solely on lexical match assumes that a term has only one meaning. While this works for pain-related terms at most times, it is important to bear in mind the issues this could cause, where specific words referenced in the text might represent something different within the phrase but are linked to

something else within SNOMED CT. For example, the clinical notes might mention a “fit note” (to provide evidence of someone’s fitness to work), but the word “fit” could be linked to the concept of seizure within SNOMED CT.

Figure 2.5. Concept normalisation for the term “pain” within SNOMED CT

This figure shows the various synonyms for the word “pain” within SNOMED CT, each of which is linked to the same SCTID. This screenshot was taken from the SNOMED CT Browser.

A review conducted by Robinson et al. (2020) looked at the use of medical ontologies in research and highlighted their benefits in addressing the challenges of heterogenous and noisy EHR narratives (Robinson & Haendel, 2020). Their compositionality may be useful in NLP because linking entities within EHR notes to concepts within SNOMED CT can disentangle some of the heterogeneity, i.e., understanding the underlying complexity and variability within clinical notes, thereby aiding better research.

UMLS is another source of structured knowledge that incorporates over a hundred biomedical terminologies, including SNOMED CT (Bodenreider, 2004). UMLS was developed to aid in effectively extracting machine-readable biomedical information by overcoming two main obstacles: the same concepts being expressed in different ways and the distribution of data across diverse, disconnected databases and systems. UMLS deals with these obstacles by

assigning each clinical term a concept unique identifier (CUI) and, similar to SNOMED CT, disambiguates them by a form of concept normalisation. For example, the concept C0009443 within UMLS is associated with the terms cold, common cold, head cold, acute rhinitis, and so on (Newman-Griffis et al., 2021). However, it lacks the compositionality that is inherent in SNOMED CT, and the difficulty of mapping terms across terminologies leads to synonymous terms occasionally being mapped to different CUIs (Fung et al., 2007). Despite this, UMLS provides a mapping from terms to concepts and includes semantic and hierarchical relationships between concepts (Cruse, 2004), therefore providing some value when integrated with the clinical text. The terms from different vocabularies that hold the same meaning are grouped together into concepts, and each concept is assigned a semantic type obtained from the UMLS Semantic Network[1], which provides a consistent categorization of all concepts and their relationships represented within UMLS. For example, “pain” has the semantic type “Sign or Symptom”, and “head” has the semantic type “Body Location or Region”. Figure 2.6 shows a portion of the Semantic Network for the semantic type “Biologic Function”. “Anxiety”, for example, would fall under the semantic subtype of “Mental Process” below.

[1] https://lhncbc.nlm.nih.gov/semanticnetwork/

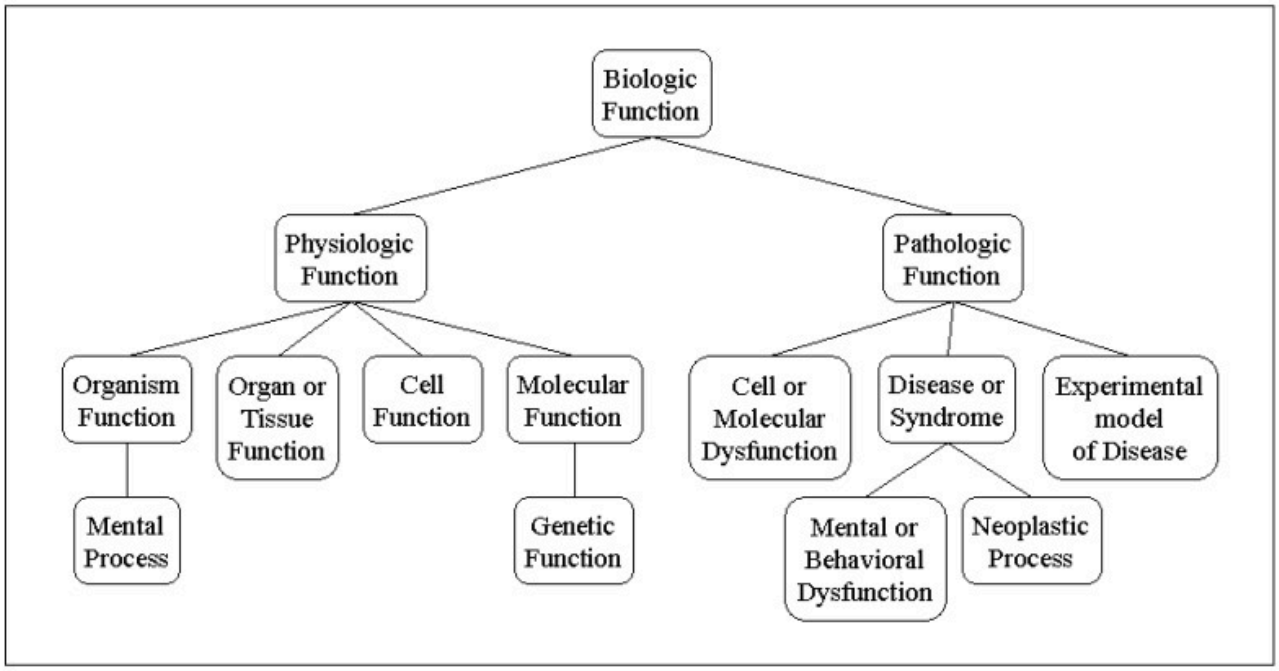


Figure 2.6. A portion of the Semantic Network for "Biologic Function"
This figure shows a part of the semantic network within UMLS.

With the increased availability of EHR data, the use of such structured knowledge has also increased, making EHR data more useable for statistical analysis and machine learning approaches (Haendel et al., 2018). The formal and hierarchical structure of such structured knowledge resources facilitates the classification of data into different taxonomic categories to combine various clinical concepts such as diseases, phenotypes, medications, and procedures (Haendel et al., 2018). They add a level of semantics to the clinical data by providing references to concepts and the relationships that exist between them, thereby enabling logical reasoning about these concepts. Combined with NLP approaches, structured knowledge can solve the problem of incorporating background knowledge (as mentioned in Section 1.1 of the Introduction chapter) and help disambiguate concepts within the clinical notes to produce more meaningful results (Haendel et al., 2018).

Such knowledge resources have also been used in combination with transformer-based architectures. Within the healthcare domain, models pre-trained on medical text from

structured knowledge resources have been shown to be beneficial for transfer learning (Walk et al., 2021). Two such models that have been developed are UmlsBERT (Michalopoulos et al., 2021) and SapBERT (Self-alignment pretraining for BERT (Liu et al., 2021)). UmlsBERT is a contextual embedding model that incorporates domain knowledge from UMLS through the use of a knowledge augmentation strategy. This approach establishes connections between words that share the same underlying concept within UMLS, as well as leveraging the knowledge of semantic types (described earlier in this section) from UMLS to generate clinically meaningful input embeddings, i.e., low-dimensional representations of words in a vector space (Mikolov et al., 2013). The former is achieved by linking words that share the same UMLS concept, thereby allowing words that have similar meanings in a medical setting to be connected, and helping the model understand that these words are related. The latter is achieved by using the semantic type information provided within UMLS, which helps the model create meaningful word representations by learning how different words fit in the broader categories of medical concepts. SapBERT follows a similar principle, where the model is pretrained on the biomedical knowledge graph of UMLS. This model utilises a self-alignment objective to cluster synonyms to the same concept (Liu et al., 2021). Self-alignment, with respect to SapBERT, refers to the model's ability to optimise its internal representational embeddings of biomedical entities. More specifically, SapBERT leverages a scalable metric learning function to generate embeddings wherein UMLS concepts with similar semantics are clustered together in the model's feature space. This metric learning loss function, a type of optimisation function, aims to improve similarity measures between data points, with the goal of ensuring that similar data points are close to each other in the model's feature space, and dissimilar ones are farther apart. The scalability of this approach stems from its ability to effectively handle large amounts of data while maintaining these relationships. Both these models can encode clinical domain knowledge into word

embeddings, which allows them to outperform existing generic and domain-specific models and produce more meaningful embeddings when compared to standard BERT models. While both models offer similar functionality, SapBERT is actively maintained and supported and has, therefore, been used in this project.

Knowledge graphs (KGs) have emerged as an efficient method of representing such conceptual data as a heterogeneous graph, with nodes representing concepts and edges representing the relationship between these concepts (Ji et al., 2022). An edge may be referred to as a predicate between subject and object nodes, together forming a subject / predicate / object triple. For example, a KG representation of concepts mentioned in the sentence, “London is located in the UK”, could model London and the UK as subject and object nodes, and location as a predicate edge, i.e., London / location / UK. KGs provide a flexible structure that aids in both reasoning with and visualisation of complex data and its interconnected relationships. This can help reveal hidden patterns and deduce new knowledge (Yoon et al., 2017).

KGs in themselves are often incomplete and cumbersome to manipulate and use (Wang et al., 2017). To overcome this, and to simplify the manipulation of KGs while preserving their structure, knowledge graph embeddings (KGEs) may be used. KGEs represent concepts and relationships within a KG as a continuous vector, or embedding, in a low-dimensional space. An example of this is shown in Figure 2.7. An embedding represents a given concept or relation in a vector space, with similar concepts and relations being represented as close to each other. Embeddings thus provide a generalised model of the KG, which can be used to infer relations not found in the original dataset.

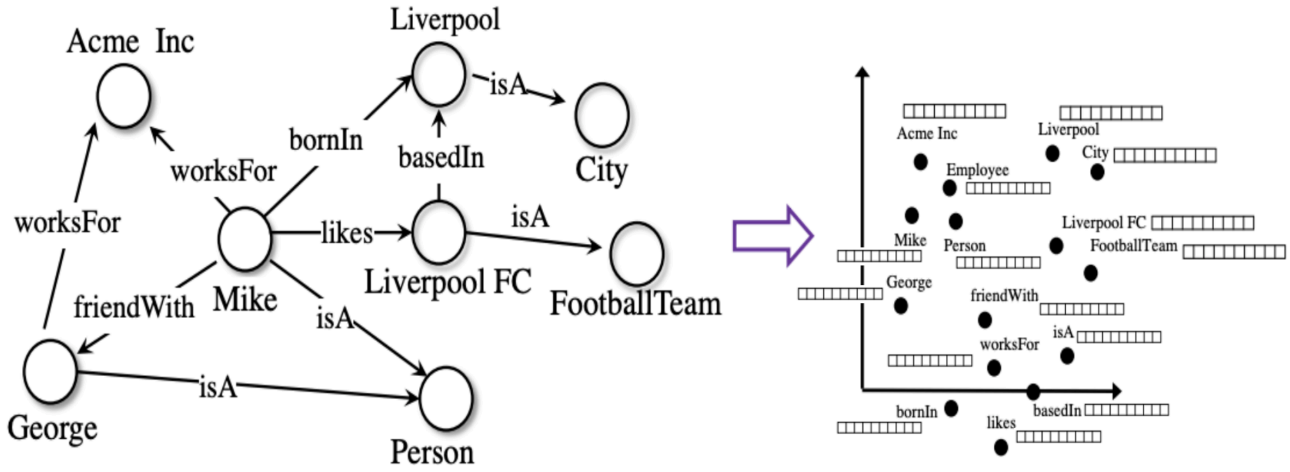


Figure 2.7. Knowledge graph represented as embeddings in a low-dimensional space (AmpliGraph, 2019b)

This figure shows an example knowledge graph on the left, with the different nodes as circles and edges as arrows. The same knowledge graph is represented as embeddings (the rectangular squares here, which in reality would be a series of numbers of fixed length) for each entity as circles, within a low-dimensional space. Here, entities that are similar to each other, such as Liverpool FC and Football Team, are closer to each other in space.

Projections of KG concepts into vector spaces as KGEs make them more scalable and can bring other latent properties to light, such as similarities between concepts (Bianchi et al., 2020). KGEs can be generated from neural models, such as TransE (Bordes et al., 2013) and ComplEx (Trouillon et al., 2016). TransE is a translation-based embedding model that uses distance-based functions to generate embeddings. The relationship between the subject (head) and object (tail) is interpreted as a translation vector (a vector that learns and stores the semantic relationship of, or "translates", two connected nodes), with the distance between the related concepts being minimised. In other words, as seen in Figure 2.8, the assumption with TransE is that the added embedding of h+r (head + relation) should be close to the embedding of t (tail) (Bordes et al., 2013).

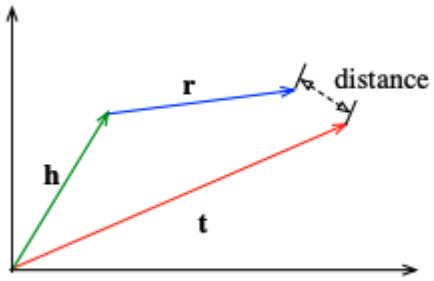


Figure 2.8. Translational distance-based embedding of TransE
(Ji et al., 2022)

This figure shows the distance between the sum of the head (h) and relation (r) being close to the tail (t), indicating related concepts since the distance is small

ComplEx, on the other hand, is a complex embedding approach that uses Hermitian dot products (complex counterparts of standard dot products) between real vectors. This embedding model is based on tensor factorisation, which is defined as the decomposition or breaking down of a tensor (a multidimensional array) into smaller, constituent parts. In this case, the embeddings are calculated by factorising (breaking down into smaller components) a three-dimensional tensor (array) in the form of n x n x m (n refers to the number of concepts - subject and object, and m refers to the number of relations or predicates) (Trouillon et al., 2016). Through tensor factorisation, this three-dimensional tensor is broken down into smaller matrices (one or two dimensions), allowing the concepts and relations to be represented by lower-dimensional dense vector embeddings, thereby reducing data dimensionality, and revealing latent patterns, i.e., underlying structures or relationships in the data that are not directly observable. Additionally, tensor factorisation allows for reconstructing the original tensor by multiplying its components back together.

KGEs have been used to represent domain knowledge from structured medical knowledge resources, and to incorporate that knowledge into NLP tasks (Chang et al., 2020). For

example, SNOMED CT has been used in several KGEs (K. Agarwal et al., 2019; Chang et al., 2020; Plaza, 2014). It comprises thousands of medical terms and their relations, organised hierarchically. Medical KGEs have been used for clustering and similarity analysis within health informatics (Chang et al., 2021; Mohamed et al., 2021), as well as inference tasks such as predicting drug targets (Fang et al., 2018; Mohamed et al., 2021; Nickel et al., 2016).

In this project, I use KGEs to examine the incorporation of domain knowledge in a text classification task and to compare classification with and without incorporated domain knowledge. Applying such methods to mental health EHRs to extract information about physical pain is novel and will yield richer information for further pain research.

2.3 Pain and EHR data

The global burden of pain is escalating, with 1 in 5 patients experiencing pain and 1 in 10 diagnosed with chronic pain annually (Goldberg & McGee, 2011). Pain is the most common reason for individuals to seek medical care (Fishman, 2007). This imposes substantial burdens across diverse populations - an estimated 25% of chronic pain stems from surgery and trauma, while 60-70% from advanced cancer and late-stage HIV/AIDS, and 70% of elderly patients experience other chronic non cancer pain (King et al., 2013). The personal and societal costs of pain is huge. In the UK alone, back pain incurs £1 billion in annual healthcare expenditures, in addition to indirect costs like absenteeism and lost productivity which is estimated between £5-10.7 billion (Maniadakis et al., 2000). Globally, the economic toll is highest on vulnerable groups including those in developing nations, the elderly, children, women, and racial/ethnic minorities (King et al., 2013). Untreated pain diminishes quality of life, which can lead to various physical, psychological, social, and financial consequences.

Inadequately managed pain can also precipitate neurological and immunological dysregulation, and increasing risk of chronic pain, if left untreated (Stephens et al., 2003). Potential consequences include depression, disability, strained relationships, suicidal ideation, and early workforce departure - the top drivers of long-term government disability benefits being musculoskeletal disorders and mental illness (Goldberg & McGee, 2011). Across countries, societies are struggling with pain's immense burden and the strains on the limited healthcare resources (Dagenais et al, 2008). The economic impact of pain exceeds most other conditions, and so understanding pain is crucial to improving public health.

Extraction of pain information from EHRs remains challenging due to the lack of standardised formats for recording such data (Fodeh et al., 2018). Without structured documentation, identifying patients experiencing pain relies heavily on clinical notes. Prior studies have utilised billing records (Pakhomov et al., 2007), structured tables or forms for pain scores (Heintzelman et al., 2013; Tian et al., 2013), and medication prescriptions (Hyun et al., 2009; Pakhomov et al., 2008) to identify pain. One such study by Tian et al. (2013) used a combination of opioid prescriptions in addition to pain diagnostic codes and pain intensity ratings within the EHR to determine pain from the records. Some studies have applied NLP methods to extract pain symptoms from such free-text clinical notes (as detailed below). However, such application remains limited compared to the utilisation of other structured EHR components in research.

While clinical notes within these records remain a relatively untapped resource for the identification of pain in patients, some research studies have applied NLP approaches to clinical notes for the extraction of pain information. For example, rules-based approaches, such as regular expressions, have shown efficacy in extracting pain information from notes (Naseri et al., 2021; Tan et al., 2018; Tian et al., 2013), and more recent machine learning

techniques have been a useful alternative method. Fodeh et al. (2018) used machine learning approaches (classifiers) on the clinical notes from the US Department of Veterans Affairs (VA) EHRs to determine pain assessment from clinical notes, where if a note had a single pain assessment subclass (determined by the authors), it was deemed as a pain-related note (Fodeh et al., 2018). Hybrid approaches combine rules-based and machine learning approaches, as seen in research by Bui et al. (2014), where they used regular expressions to determine pain within clinical notes from the US VA EHRs to determine whether snippets of text mentioned pain or not. Over 25,000 regular expressions were created on their pain dataset by utilising a regular expression discovery (RED) algorithm, which was then combined with classifier models to achieve an accuracy of 81% (Bui & Zeng-Treitler, 2014).

Beyond identification, some studies extract indicators of care quality. Dorflinger et al. (2014) developed an information extraction tool for extracting information about the quality of pain care in a primary care setting at a VA healthcare facility, which included information about the quality and documentation of pain assessment, pain treatment, and reassessment during primary care appointments. They used indicators such as explicit mentions of the word “pain”, pain-related assessments such as pain-related conditions being mentioned in X-ray or MRI reports, pain-related medications or referrals and other pain management measures like exercise and pain education. Their final tool consisted of 12 dichotomously scored indicators assessing the quality and documentation of pain care in three domains: assessment, treatment, and reassessment (Dorflinger et al., 2014).

Additionally, some research has been conducted on specific subtypes of pain. A systematic review by Bacco et al. (2022) looked at the NLP approaches applied to the identification of low back pain and spine diseases. They identified 16 relevant papers and found that the research in these papers used either rules-based approaches, such as regular expressions,

or machine learning based approaches, focusing on radiology images in combination with clinical notes, using methods such as logistic regression and BERT (Bacco et al., 2022). Furthermore, Vandenbussche et al. (2022) used NLP on self-reported narratives from patients on their migraine or cluster headache experiences at a hospital in a Dutch-speaking part of Belgium. Their analysis included thematic and sentiment analysis, which revealed largely negative sentiments in texts by patients with migraine and/or cluster headaches. They also applied logistic regression and support vector machine algorithms, achieving F1 scores for detecting cluster headaches between 0.82 and 0.86 (Vandenbussche et al., 2022). Similarly, Nunes et al. (2022) conducted unsupervised topic modelling to recognise complex patterns in spontaneous verbal descriptions of chronic pain and use these patterns to quantify and qualify experiences of pain (Nunes et al., 2021).

As evidenced by these studies, both rules-based and machine learning based NLP approaches show potential for the identification and extraction of pain information from clinical notes. However, most existing studies still rely on structured fields within the EHR databases to some extent and are limited to extracting pain assessment or care quality information. Pain remains poorly coded within EHRs, and there is a need for exploring other methods to identify pain-related records. Additionally, there is a lack of research in the study of pain in mental health settings. As highlighted from the studies mentioned at the beginning of this section (Section 2.3), the link between mental health and pain is very important, yet understudied. Individuals with mental health disorders experience a disproportionately high burden of pain, which can exacerbate their mental health symptoms and delay recovery. Conversely, pain can contribute to the development or worsening of mental health conditions like depression and anxiety. There is also a lack of the application of NLP methods in this area through the incorporation of any external domain knowledge. This thesis addresses these gaps by developing NLP methods that integrate external domain knowledge and apply them to the

extraction of pain information in a mental health setting. This is novel and essential for the advancement of pain research in the complex mental health domain.

CHAPTER 3: Development of a Pain Lexicon

3.1 Foreword

The preceding chapters discuss the uncertainty and ambiguity surrounding how pain is mentioned in clinical notes. Given the various ways pain can be mentioned (for example, ache, sore, myalgia), identifying documents discussing pain becomes quite challenging. This challenge prompted the need for a comprehensive dictionary or lexicon containing a wide range of pain-related terms, covering the diverse expressions of pain within clinical notes. Notably, existing literature and ontology databases did not contain such a comprehensive lexicon. Consequently, a decision was made to construct such a lexicon from scratch. This chapter describes the development of such a lexicon of pain terms, for use in retrieval of pain-related texts, that can be used in a downstream classification task. In addition to literature and vocabulary-based sources of pain terms, the lexicon also utilises social media sources such as Twitter and Reddit to capture the patient's voice. The developed lexicon was shared with a PPI group to gather input on whether the terms captured the way in which service users and carers would describe their pain. The lexicon has been made available under an open-source license to the general research community for use in other pain-related research and will be formalised as an ontology as part of future work.

This work was presented at the Health Text Analytics Conference 2021 (HealTAC 2021) and subsequently published in the journal *Frontiers in Digital Health*.

Chaturvedi J, Mascio A, Velupillai SU, Roberts A*. Development of a Lexicon for Pain.*

Frontiers in Digital Health. 2021; 3: 193. Available from:

https://www.frontiersin.org/article/10.3389/fdgth.2021.778305

I contributed as the first author, designed the research, and drafted the manuscript. The following sub-sections of this chapter reproduce the published paper, with some minor formatting adjustments to keep it in line with the thesis format. The content itself has not been altered.

Some of the code for building embedding models, as well as one of the embedding models used in this work, were developed by Dr Aurelie Mascio (Mascio, 2022).

Section 3.11, which follows the journal article, is not part of this publication and details the role of a PPI group in validating the lexicon.

Development of a Lexicon for Pain

Jaya Chaturvedi[1]*, Aurelie Mascio[1], Sumithra Velupillai[1], Angus Roberts[1,2]
[1]Institute of Psychiatry, Psychology and Neurosciences, Department of Biostatistics and Health Informatics, King's College London, London, United Kingdom
[2]Health Data Research UK, London, United Kingdom

Corresponding Author
jaya.1.chaturvedi@kcl.ac.uk

Keywords: lexicon[1], pain[2], natural language processing[3], electronic health record[4], mental health[5].

3.2 Abstract

Pain has been an area of growing interest in the past decade and is known to be associated with mental health issues. Due to the ambiguous nature of how pain is described in text, it presents a unique natural language processing (NLP) challenge. Understanding how pain is described in text and utilizing this knowledge to improve NLP tasks would be of substantial clinical importance. Not much work has previously been done in this space. For this reason, and in order to develop an English lexicon for use in NLP applications, an exploration of pain concepts within free text was conducted. The exploratory text sources included two hospital databases, a social media platform (Twitter), and an online community (Reddit). This exploration helped select appropriate sources and inform the construction of a pain lexicon. The terms within the final lexicon were derived from three sources—literature, ontologies, and word embedding models. This lexicon was validated by two clinicians as well as compared to an existing 26-term pain sub-ontology and MeSH (Medical Subject Headings) terms. The final validated lexicon consists of 382 terms and will be used in downstream NLP tasks by helping select appropriate pain-related documents from electronic health record (EHR) databases, as well as pre-annotating these words to help in the development of an NLP application for the

classification of mentions of pain within the documents. The lexicon and the code used to generate the embedding models have been made publicly available.

3.3 Introduction

Pain is known to have a strong relationship with emotions, which can lead to damaging consequences (Heintzelman et al., 2013). This is worsened for people suffering from persistent pain. It can lead to long-term mental health effects such as the "secondary pain effect", which encapsulates the strong feelings toward the long-term implications of suffering from pain (Heintzelman et al., 2013). The Biopsychosocial framework of pain reiterates the multidimensionality of pain and explains the dynamic relationships of pain with biological, psychological, and social factors (Merlin et al., 2014). Pain has been an active area of research, especially since the onset of the crisis of opioid use in the United States (Howard et al., 2018). Pain also has a significant impact on the healthcare system and society in terms of costs (Groenewald et al., 2014). Apart from research, it has also been of interest to the general population. Figure 2 shows Google trends for the search term "pain" over time (2004 to present) compared with two other common symptoms ("fever" and "cough") to investigate whether the trends are reflective of a general increase in searches, or an actual increase in search of the term. All three terms were selected as "medical terms" rather than "general search" terms to avoid any metaphorical mentions and make the words more accurately comparable. This was possible through the use of a Google Trends[2] feature which allows the user to choose the search category (generic "Search term" category would include any search results for the word "pain," "Medical condition"/"Disease" category would only include "pain"

[2] https://trends.google.com/trends/explore?date=all&q=%2Fm%2F062t2

when searched as a medical condition or disease). Pain shows an incremental increase worldwide (Figure 3.1).

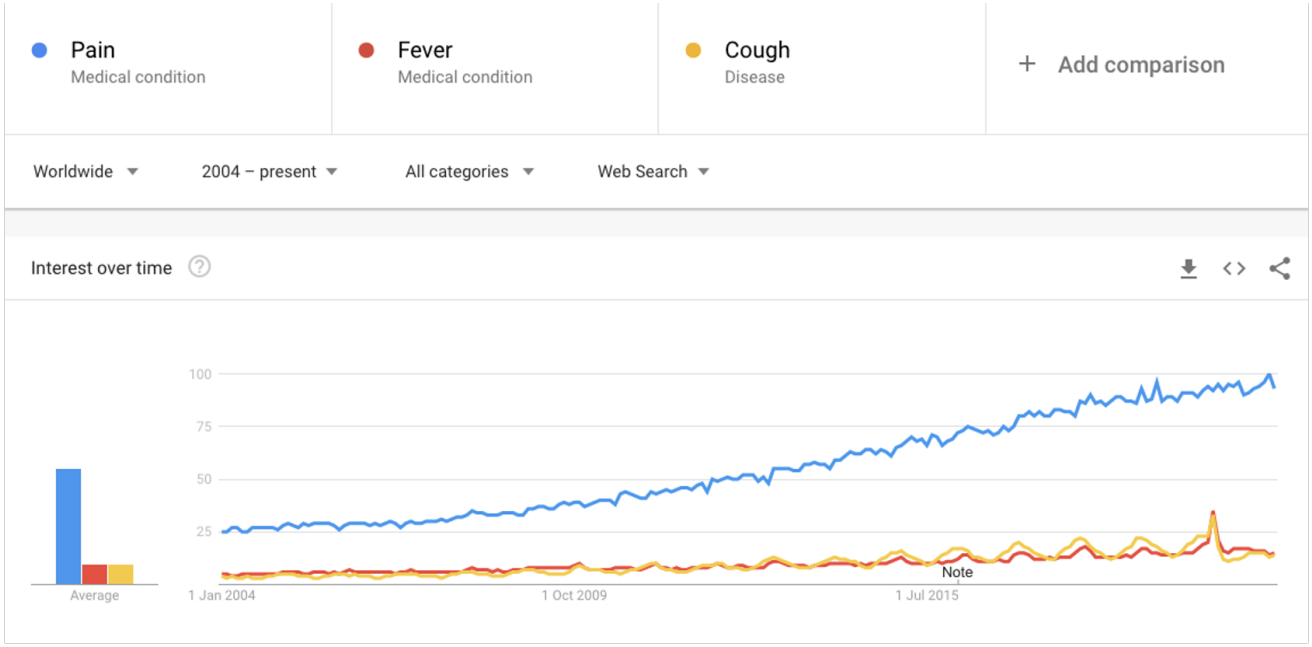


Figure 3.1. Google trends for medical condition search term "pain" compared to other common symptoms "fever" and "cough."
X-axis represents time in years. Y-axis numbers represent the search interest relative to the highest point on the chart (100 is the peak popularity for the term, 50 indicates the term is half as popular, and 0 means there was insufficient data for the term).

Research is a growing secondary use of mental health electronic health records (EHRs), specifically the free-text fields (Stewart et al., 2009). It has the potential to provide additional information on contextual factors around the patient (Velupillai et al., 2018). While it is beneficial to include clinical notes in research, extracting, and understanding information from the free text can be challenging (Mascio et al., 2020). NLP methods can help combat some of the issues inherent in clinical text, such as misspellings, abbreviations, and semantic ambiguities.

Another rich source of health-related textual data is social media, as it provides a unique patient perspective on health (Foufi et al., 2019). In recent years, there has been an increase in the use of social media platforms to share health information, receive and provide support, and look for advice from others suffering with similar ailments (Foufi et al., 2019). Content from these platforms has also been increasingly used in health research. Examples include finding symptom clusters for breast cancer (Marshall et al., 2016), understanding the relationships between e-cigarettes and mental illness (Sharma et al., 2016), as well as understanding user-generated discourse around obesity (Chou et al., 2014). The main platforms involved in these studies have been Reddit[3] and Twitter[4]. Reddit has been a good source for such textual research due to its wide usage as well as the ability to post anonymously (Johnson & Ambrose, 2006). Reddit has more than 330 million active monthly users, and over 138K active communities (Social Media Today, 2018). A key feature of Reddit is the subforum function which allows creation of subreddit communities dedicated to shared interests (Foufi et al., 2019). Twitter has shorter text spans than Reddit, a maximum of 280 characters (Boot et al., 2019). Despite this limitation, Twitter is widely used in research around mental health and suicidality (Choudhury et al., 2021; Coppersmith et al., 2015; De Choudhury et al., 2016).

The term “pain” presents a unique NLP problem, due to its subjective nature and ambiguous description. Pain can refer to physical distress, or existential suffering, and sometimes even legal punishment (Carlson & Hooten, 2020). However, within the clinical context, it will most likely be the former two. It also has metaphorical uses in phrases such as “for being a pain” (Carlson & Hooten, 2020). In order to better understand how pain is described in different textual sources, and to construct a lexicon of pain for use in NLP applications, this study does

[3] https://www.reddit.com/
[4] https://twitter.com/home?lang=en

a preliminary exploration of mentions of pain. This exploration includes analysis of mentions of pain in four different sources with the objective of understanding how mentions of pain differ in these sources, and whether they cover common themes. These exploratory sources include—a mental health hospital in the UK (CRIS, from the South London and Maudsley NHS Foundation Trust), the critical care units of a hospital in the United States (MIMIC-III), Reddit, and Twitter.

Gaining a good understanding of how pain is mentioned in text can be formalised by creation of a lexicon of pain terms. Lexicons are a valuable resource that can help develop NLP systems and improve extraction of concepts of interest from clinical text (Velupillai et al., 2016). Lexicons provide a wide range of terms and misspellings from relevant domains, which will be advantageous in future NLP tasks and will minimise the risk of missing important documents that contain these relevant terms. An existing ontology, The Experimental Factor Ontology (Koscielny et al., 2021), consists of a subsection of 26 pain related terms, but to our knowledge, no previous studies have explored how the concept of pain is used in different text sources, and used this to generate a new lexicon. While using terms generated by a domain expert has the benefit of being more precise, we believe that for an ambiguous term such as pain, our method of producing a lexicon semi-automatically, for domain expert review, will favour recall without damaging precision (i.e., sensitivity without loss of positive predictive value). The generation of this lexicon will involve a combination of terms related to pain from three sources—literature, ontologies, and embedding models built using EHR data. Mentions from social media that were part of the exploratory sources are not included as lexicon sources since the primary purpose of the lexicon in this instance is for use on EHR data. Any relevant mentions from social media may be added to the lexicon at a later date.

The aim of this study was to conduct an exploration of how pain was mentioned within four different text sources. The purpose of this exploration was to understand what sources of textual information might be useful additions to the lexicon. The eventual goal of generating this lexicon is to be able to use it in downstream NLP tasks where it can be used to identify relevant pain-related documents from EHR databases.

3.4 Materials and Methods

The final lexicon consists of relevant pain related terms from three key areas—ontologies, literature, and embedding models. The lexicon was reviewed and validated by domain experts. In addition to this, the lexicon was also compared to another ontology that consists of 26 pain-related terms. This ontology is available as part of the Experimental Factor Ontology (version 1.4) (Koscielny et al., 2021) as a subsection for pain.

3.4.1 Data Collection and Exploration/Source Comparison

Four different data sources were explored for mentions of pain within their textual components, and a comparison was conducted to understand the different contexts in which pain can be mentioned. Fifty randomly selected documents were extracted from each source. The number of documents was limited to 50 per text source for pragmatic reasons: manual review is a labour-intensive process. This decision should not impact the lexicon development, as these documents are used only for exploration, with embeddings built on the whole of two sources (MIMIC and CRIS) were used to generate the terms for the lexicon to supplement the development of the lexicon.

Ethics and Data Access

While data from Reddit and Twitter are publicly available, applicable ethical research protocols proposed by Benton et al. were followed in this study (Benton et al., 2017). No identifiable user data or private accounts were used, and any sensitive direct quotes were paraphrased.

Data from Twitter is available through their API after approval of registration for access to this data, details of which can be found in their general guidelines and policies documentation (Twitter Help Center, n.d.). Data access information for CRIS (NIHR Maudsley Biomedical Research Centre, n.d.) and MIMIC-III (Johnson, Alistair et al., 2015) are detailed on their respective websites.

CRIS

An anonymised version of EHR data from The South London and Maudsley NHS Foundation Trust (SLaM) is stored in the Clinical Record Interactive Search (CRIS) database (Stewart et al., 2009). The infrastructure of CRIS has been described in detail (Perera et al., 2016) with an overview of the cohort profile. This project was approved by the CRIS oversight committee (Oxford C Research Ethics Committee, reference 18/SC/0372). Clinical Record Interactive Search consists of almost 30 million notes and correspondence letters, with an average of 90 documents per patient (Velupillai et al., 2018).

A SQL query was run on the most common source of clinical text (“attachments” table which consists of documents such as discharge and assessment documents, GP letters, review, and referral forms) within the CRIS database, and 50 randomly selected documents that contained the keyword “pain” (both upper and lower case) were extracted. This would include any instance of “pain” regardless of whether it refers to physical pain or emotional/mental

pain. Other features of the documents, such as maximum and minimum length of documents were calculated, as well as common collocates for the term “pain”.

MIMIC-III

Medical Information Mart for Intensive Care (MIMIC-III) is an EHR database which was developed by the Massachusetts Institute of Technology (MIT), available for researchers under a specified governance model (Johnson, Alistair et al., 2015). Medical Information Mart for Intensive Care consists of about 1.2 million clinical notes (Nuthakki et al., 2019).

A SQL query was run on the “note-events” table which contains majority of the clinical notes (such as nursing and physician notes, ECG reports, radiology reports, and discharge summaries) within the database, and 50 random documents containing the keyword “pain” (both upper and lower case) were extracted. Like the CRIS database, an analysis of the maximum and minimum length of documents was carried out, and common collocates for the term “pain” were explored.

Reddit

Reddit is an online community which supports unidentifiable accounts to allow users to post anonymously and provides sub communities for people to discuss topics of shared interest. The chronic pain subreddit (r/ChronicPain) community was used in this study. Other subreddits around pain included more specific communities, such as “back pain,” which would not serve our purpose of keeping it general. While this approach might miss mentions of other types of pain, there didn’t seem to be a way around this due to absence of a general pain subreddit. Data from Reddit was extracted using the python package PRAW (Boe, 2012). No time filter was applied. Seven thousand seven hundred posts were extracted, out of which 50 posts were randomly selected.

Twitter

Twitter is an online micro-blogging platform with an enormous number of users who post short (280 characters or less) messages, referred to as “tweets,” on topics of interest. It is a good resource for textual data because of the volume of tweets posted on it and the public availability of this data (Bian et al., 2012). Python package tweepy (Roesslein, 2020) was used to extract tweets using the search term “chronic pain”. As with Reddit, chronic pain was used instead of pain to help get more meaningful health-related results. This approach was not applied to the EHR text as the assumption was that metaphorical mentions would be more prevalent in social media. This does carry the risk of possibly missing out on mentions of pain that were not explicitly chronic. Since the Twitter API allows[5] for extraction of tweets within a seven day window, 7,707 tweets were extracted within the time period 06/08/2020 to 11/08/2020 that consisted of the keywords “chronic pain” (case insensitive). Out of these, 50 tweets were randomly selected for analysis.

3.4.2 Lexicon Development

Concordances and analyses on data from the previous step were used to inform the appropriateness of the mentions of “pain” and whether they were meaningful mentions and thereby suitable for inclusion in building a lexicon of pain terms. The terms within the EHR text had more appropriate concordances (i.e., referring to actual pain rather than metaphorical mentions) and were therefore included in the lexicon while the social media ones were not. Embedding models built using Twitter (Pennington et al., 2014) and Reddit (A. Agarwal, 2015) data were not used as their results returned words that did not seem relevant to the term “pain.” They generated terms such as brain, anger, patience, and habit with Twitter, and

[5] X (formerly Twitter) data is not available for free anymore. This work was conducted before it was moved behind a paywall.

words such as apartment, principal, and goal by Reddit. In addition to this, a few publications and ontologies were explored as potential sources as well. The final lexicon was built by combining terms generated through three different sources.

Literature-Based Terms

We harvested pain-related words from three publications:

(1) A list of symptom terms provided by a systematic review on application of NLP methods for symptom extraction from electronic patient-authored text (ePAT) (Dreisbach et al., 2019). Some examples include pain, ache, sore, tenderness, head discomfort.

(2) Ten words most similar to pain generated in a survey of biomedical literature-based word embedding models (Khattak et al., 2019). Some examples include discomfort, fatigue, pains, headache, backache.

(3) A list of sign and symptom strings generated using NLP to meaningfully depict experiences of pain in patients with metastatic prostate cancer, as well as identify novel pain phenotypes (Heintzelman et al., 2013). In our literature search, this was the only paper on NLP-based extraction of pain terms that included a list of the terms used. Some examples include ache, abdomen pain, backpain, arthralgia, bellyache.

These lists were cleaned by lowercasing all terms, and only keeping terms made up of one or two tokens as these included most of the terms, and any terms with more than two tokens were less meaningful or repetitive of the two token terms. Terms with more than two tokens were only listed in one of the papers (Heintzelman et al., 2013), and some examples of these were terms such as pain of jaw, right lower quadrant abdominal pain, upper chest pain, and so on, most of which were covered within the two token terms such as abdominal pain and chest pain.

Ontology-Based Terms

We incorporated synonyms for pain from three biomedical ontologies—The Unified Medical Language System (UMLS) (Bodenreider, 2004), Systematized NOmenclature of MEDicine Clinical Terms (SNOMED CT) (Stearns et al., 2001), and International statistical Classification of Diseases and related health problems: tenth revision (ICD- 10) (World Health Organization, 2004). Unified Medical Language System contains concepts from SNOMED CT and ICD-10, in addition to several other vocabularies. From each, we extracted terms of up to two tokens that either matched "pain∗," were synonyms of pain, or described as child nodes of pain.

Embedding Models

Embedding models (Mikolov et al., 2013; Y. Wang, Liu, et al., 2018) using eight different parameters and four different text sources were used to generate additional words similar to "pain." The elbow method (Ye & Fabbri, 2018) was used to determine the cut-off point in word similarity which helped determine the similarity threshold for each model. An advantage of using embedding models is their ability to capture misspellings. Any duplicates were removed, and the remaining terms were added to the lexicon.

Two of the embedding models [both described in Viani et al. (Viani et al., 2019)] were built using clinical text available within the MIMIC-II database (Saeed et al., 2011). Four embedding models were built using clinical text available within MIMIC-III, of which three were built using genism implementation of word2vec (Mikolov et al., 2013; Y. Wang, Liu, et al., 2018)) and one using FastText (Bojanowski et al., 2017). One model was built using word2vec over a severe mental illness (SMI) cohort from CRIS. Finally, a publicly available model built on PubMed and PubMed Central (PMC) article texts was used (Pyysalo et al.,

2013). Only unigrams were included from all the models. The parameters for these are detailed in Table 6.

3.4.3 Validation

Upon collection of data from the four different sources, common themes were explored. The purpose was to understand the common contexts in which pain might be mentioned. In addition to common themes, length of the text containing mentions of pain was calculated, along with most frequent concordances and mutual information scores.

Validation of the terms for inclusion in the final lexicon was conducted using two methods—validation by two clinicians, comparison to an existing pain-related lexicon, and comparison to MeSH[6] (Medical Subject Headings).

A list of the terms generated through the three text sources was shared with two clinicians who marked each term as: relevant mention of pain, not relevant to pain, or too vague in relation to pain. In addition to this, they added a few new terms to the lexicon.

As an additional validation step, the final lexicon validated by the clinicians was compared to an existing ontology, The Experimental Factor Ontology (Koscielny et al., 2021), which consists of a sub-section of 26 pain-related terms. The final lexicon was also compared to 63 pain-related MeSH terms. Each MeSH term also consisted of a set of entry terms (a total of 941 pain-related terms). Entry terms refer to synonyms, alternate forms, and other terms that are closely related to the MeSH term (National Library of Medicine, 2021). With both these comparisons, any terms that did not overlap were investigated to see why they might be missing from our lexicon.

[6] https://www.ncbi.nlm.nih.gov/mesh/?term=pain

After generation and validation of the final lexicon, the pain- related terms were separated out from the terms (such as pain from leg pain, arm pain; sore from sore mouth, sore muscle, etc.) and these terms were looked up within a cohort of SMI patients from the CRIS database. A frequency count was conducted to see which of these terms occur most frequently within this cohort of patients.

3.5 Results

3.5.1 Exploration of Pain

Three common pain terms were chosen to gain an understanding of how frequently they are mentioned in EHR documents. These terms were: pain, chronic pain, and words ending with -algia, a common suffix meaning pain. A more detailed search on other pain-related terms such as ache will be conducted at a later stage. A summary of frequencies of these terms within the two EHR- based sources is outlined in Table 3.1. As seen in the table, the term “pain” had the greatest number of mentions and was thus used for selecting documents from the databases for exploration (as described in Section 3.4).

Comparing the EHR text data to those from social media platforms Twitter and Reddit, the length of text containing the word “pain” was calculated to understand how much content might be available in each source (Table 3.2).

During the comparison of these sources, four common themes emerged, as shown in Table 3.3.

An analysis was conducted using Lancsbox (Brezina et al., 2020) to get the collocates associated with the term “pain,” limiting to only those words that had a frequency of more than 10. The top five collocates from the different sources are listed in Table 3.4. Reddit and Twitter produced mostly generic terms which were not very meaningful.

The collocation tool within LancsBox looks at five words on either side of the search term “pain,” which explains why “pain” is also a collocate within the Reddit dataset since there were instances of mentions of “pain” as can be seen in these paraphrased examples— “I suffer from a condition which causes back pain and pain in legs”; “I have chronic pain. The pain is in my shoulder...” and could also be why generic words like “anyone” (instances such as “I have tried opioids for back pain. Has anyone else seen an improvement with this...”; “Has anyone used heat for pain...”) and “anything” (instances such as “the meds are not doing anything for my pain”) have been selected.

Table 3.5 lists out the top five collocates for “pain” with a mutual information (MI) score >6. MI score measures the amount of non-randomness present when two words occur (Hunston, 2002) thereby giving a more accurate idea of the relationship between two words (Smyth, 2010). It is recommended that an MI score greater than 3 be used (Smyth, 2010) to get more meaningful results. An MI score of 5 and more was used in this instance since collocates with a lower MI score were generic and vague, including words such as “what,” “if,” and “with.” The letters in the brackets indicate whether they occurred to the right (R) or left (L) of the word “pain.” Reddit and Twitter data produced mostly generic results.

Using the observations made during this preliminary exploration, a conceptual diagram (Figure 3.2) of pain was created. The objective of constructing this conceptual diagram was to visualise what features were commonly found around the mention of pain.

Terms	CRIS - Attachments	MIMIC-III
pain	29.59	44.13
chronic pain	1.22	4.04
*algia	1.14	1.44

Table 3.1. Count of mentions of "pain", "chronic pain", and "-algia" per 10,000 tokens within the two databases – CRIS and MIMIC-III

(counts for "pain" include "chronic pain" instances too)

Source	CRIS	MIMIC	Twitter	Reddit
Average length of text (charac.)	8,144	3,864	62	1,065
Minimum length of text (charac.)	1,155	165	11	139
Maximum length of text (charac.)	32,767	9,549	106	3,598

Table 3.2. Length of text within documents containing the word "pain" in the 4 text sources on a random set of 50 documents for each text source

Source	Quality/type of pain	Feelings/experiences associated with the pain	Medication or other measures	Related to body parts
CRIS	…in constant pain.. …ongoing pain… …pain was quite severe..	…overwhelmed by chronic pain problems… …fear of pain… …pain causing distress.. …struggles with chronic pain…	…drugs to numb the pain… …pain relief medication not controlling the pain… …side effects from pain relief medication… …no pain relief with NSAIDs…	…chronic back pain.. …chest pain…
MIMIC-III	…severe pain… …atypical pain…	-	…PO as needed for pain… …taking narcotic pain medication… …managed with IV pain medication…	…chronic back pain… …chest pain… …abdominal pain… …right leg pain…

			…and pain was controlled with oral analgesics..	…chronic lower back pain…
Reddit	Sharp pain.. Widespread pain..	…could be causing pain …painful trips to the kitchen … in the same painful position as 3 months ago…	…helped my back pain…	Shoulder pain Back pain Chronic neck pain Chronic joint pain
Twitter		…to live pain-free	…muscle painbuster ..	…Joint muscle pain …Back pain

Table 3.3. Common themes around "pain" in the 50 randomly selected documents from the four data sources with examples for each

CRIS	**MIMIC-III**	**Reddit**	**Twitter**
chronic	control	pain	agony
back	acute	about	amazingly
clinic	chronic	anyone	achieved

physical	assessment	back	american
health	plan	anything	body

Table 3.4. Collocates for "pain" with frequency >10

Collocates refers to terms that occur around the word of interest - pain

CRIS	**MIMIC-III**	**Reddit**	**Twitter**
killers (R)	chronic (R)	board (R)	people (L)
chronic (L)	control (L)	certified (L)	amp (R)
fibromyalgia (R)	complains (L)	suboxone (L)	get (L)
ongoing (R)	incisional (L)	chronic (L)	medical (L)
feet (R)	acute (L)	doctor (R)	suffer (L)

Table 3.5. Collocates for "pain" with an MI score > 6

MI score refers to a measurement of the amount of non-randomness present when two words co-occur

3.5.2 Building the Lexicon

Table 3.6 summarises the number of words obtained from the three different sources. For the embedding models, the model parameters and elbow thresholds are also included.

After compiling the words from all these sources, the total size of the lexicon was 935 words (including duplicates and 57 misspellings), with 35% of them being unigrams and 65% bigrams. The most frequently occurring words in the final lexicon were pain (n = 46), discomfort (n = 10), headache (n = 8), soreness (n = 8), and pains/painful/ache/backache (n = 7). Table 3.7 shows the coverage of the lexicon at this stage.

The Venn diagrams of the unique terms are shown in Figure 3.3. A total of six terms overlap between the three sources, with the most overlap (54 terms) being between literature and ontology. There is no overlap between all three ontologies, with the most overlap (27 terms) being between SNOMED CT and UMLS. There is no overlap between ICD-10 and UMLS due to the former consisting of mostly three-token terms, while the terms in all sources have been limited to up to two tokens. For example, ICD-10 consists of terms such as pain in limb, pain in throat, pain in joints, rather than limb pain, throat pain, and joint pain. There was no overlap at all between the different embedding models in Figure 3C, and the inclusion of all 3 models within the lexicon indicates wide coverage of terms that are used in these differing sources. A comparison of the two MIMIC models (MIMIC-II and MIMIC-III) showed that they generated unique terms with minimal overlap, thereby justifying the use of both versions. Apart from the overlap between common base words such as pain and cramps, the differences were mainly in misspellings, as well as MIMIC-II containing a large number of concatenated words such as painburning and paintenderness, which was not the case in MIMIC-III. This could be due to different data formats used in both versions – MIMIC-II notes were in CLOB format (stores data in random-access chunks) while the text within MIMIC-III was in string format.

After post-processing to remove duplicates, punctuations/symbols, and words of less than four characters, the lexicon was validated by two clinicians, leading to a final size of 382 terms (Figure 3.4).

Source	**Parameters**	**Elbow threshold**	**No. of unigrams**	**No. of bigrams**	**Total no. of words**
Literature	-	-	71	170	241
Ontologies			83	440	523

UMLS	-	-	11	70	81
SNOMED CT	-	-	67	368	435
ICD-10	-	-	5	2	7
Embedding models			171	-	171
MIMIC-II	w2v, size=100, window=5, min_count=15, workers=4	0.57	33	-	33
MIMIC-II	w2v, size=400, window=5, min_count=15, workers=4	0.47	40	-	40
MIMIC-III	w2v, size=100, window=5, min_count=15, workers=4	0.66	4	-	4
MIMIC-III	w2v, size=400, window=5, min_count=15, workers=4	0.47	12	-	12

MIMIC-III	w2v, size=300, window=10, min_count=5, workers=16	0.44	26	-	26
MIMIC-III	FastText, size=300, window=10, min_count=5	0.93	30	-	30
CRIS (SMI)	w2v, size=300, window=10, min_count=5	0.69	16	-	16
PubMed	w2v, size=200, window=5	0.73	10	-	10

Table 3.6. Number of words obtained from the different sources, and parameters/elbow threshold for the embedding models

w2v (word2vec)or FastText refers to the algorithm used; size refers to the dimensionality of the word vectors i.e., each word will be represented as a vector of (100/200/300/400) dimensions; window determines the size of the context window for the model to consider when training; min_count specifies the minimum number of occurrences a word must have in the corpus to be included in the vocabulary; workers specifies the number of worker threads to use for training the model, which will be responsible for parallelising and accelerating the training process by distributing the workload across multiple CPU cores.

Lexicon source	**# of unique terms**	**Total # of terms**
Literature	218	241
Ontologies	291	523

Embeddings	68	171

Table 3.7. Lexicon coverage

This table shows the distribution of terms across the 3 sources used to build the lexicon.

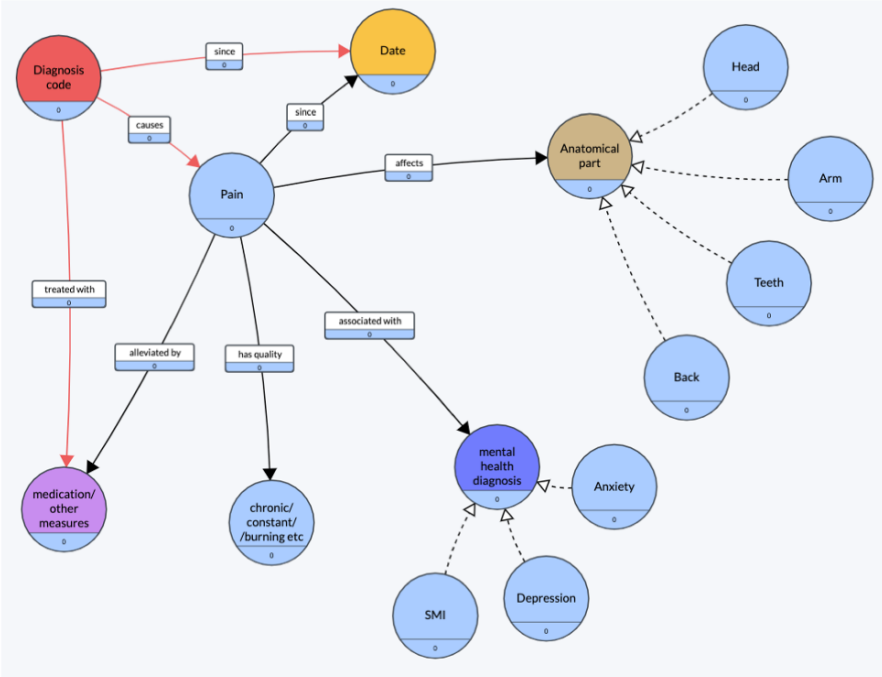


Figure 3.2. Conceptual diagram of pain
Created using an online tool, Grafo (Gra.fo, 2020)

This conceptual diagram shows the various relations of pain, from diagnosis code causing the pain, to various mental health diagnoses associated with the pain. The circles represent the concepts and the rectangles within the arrows indicate the relationships between the concepts. The colours of the circles are arbitrary and do not hold any meaning.

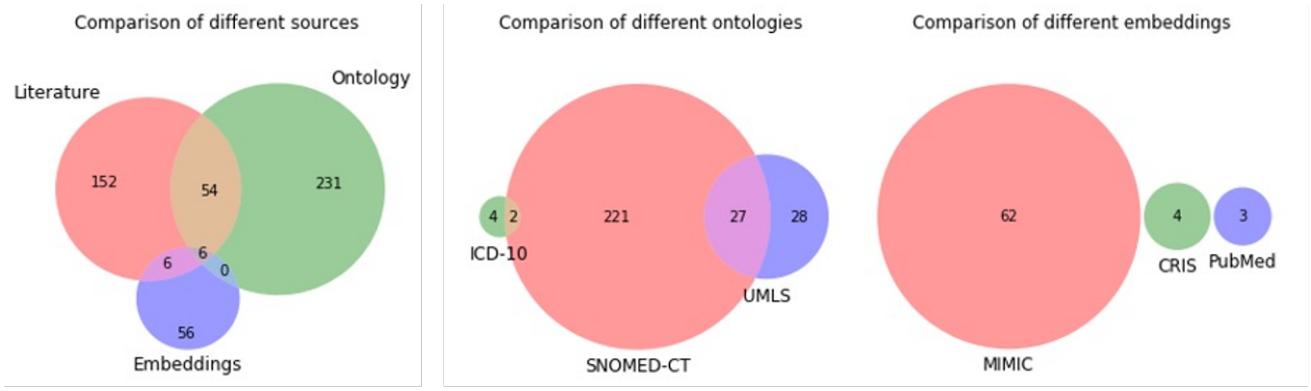


Figure 3.3. Venn diagram of unique terms generated from the different sources (A), different ontologies (B), and different embedding models (C).

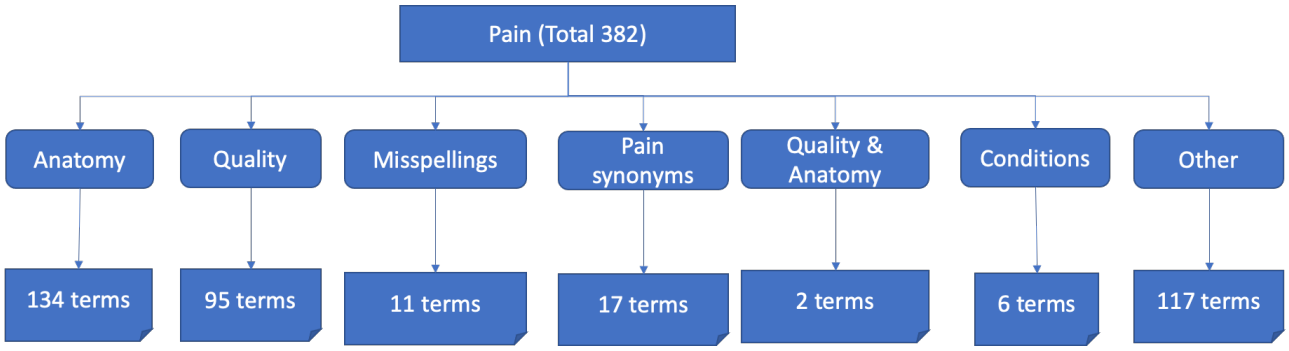


Figure 3.4. Distribution of terms within pain lexicon.
The terms within the lexicon fell into 7 different categories which are displayed in this diagram.

The final pain lexicon and the code to generate the embedding models is openly available on GitHub[7] and will also be added to other ontology collections such as BioPortal[8].

Some patterns were identified within the lexicon which enabled generation of a shorter list of pain terms which captured all the other terms within the patterns, such as the word “pain” capturing “chest pain,” “burning pain,” and ache capturing “headache,” “belly ache,” etc. For

[7] https://github.com/jayachaturvedi/pain_lexicon
[8] https://bioportal.bioontology.org/

example, terms such as “chest pain,” “head discomfort,” “aching muscles,” follow a pattern of <anatomy> followed by <pain term> or vice versa; terms like “burning pain” and “chronic pain” follow a pattern of <quality term><pain term>, and some are a combination of quality and anatomy such as “chronic back pain” which follows a pattern of <quality term><anatomy term><pain term>.

A frequency count of some other common pain related terms [using wildcard character (%) to capture any words containing these terms] was conducted on a cohort of SMI patients within the CRIS EHR documents. Top 13 terms are listed out in Table 3.8.

Keyword	**Percentage (Over entire cohort)**
%ache%	54%
%pain%	36%
%burn%	7%
%sore%	3%
%algia%	< 1%
%spasm%	< 1%
%dynia%	< 1%
%algesia	< 1%
colic%	< 1%
hurt%	< 1%

sciatic%	< 1%
tender%	< 1%
cramp%	< 1%

Table 3.8. Top 13 common pain-related keyword terms within a cohort of patients (n=57,008) in the CRIS database

3.5.3 Validation of the Lexicon

Two forms of validation were carried out on the lexicon— validation by two clinicians, and validation against an existing ontology of pain terms.

Upon validation by the clinicians, 11 new terms were added to the lexicon and 39 terms were removed from the lexicon. The reasons for removal of words were when they were too ambiguous and non-specific (such as fatigue and complaints), and words that did not indicate pain per se (such as itchiness, nausea, paraesthesia, tightness). Some examples of terms that were removed are algophobia, bloating, fatigue, and nausea. Terms added were acronyms (such as LBP for lower back pain), pain education, antalgic gait.

The Experimental Factor Ontology (Koscielny et al., 2021) contains a pain sub-section consisting of 26 pain related terms. Upon comparison with our lexicon, it was found that 18 (69%) of the terms within the Experimental Factor Ontology matched. Amongst the ones that did not match, most were words with three tokens, which would have been excluded from our lexicon. They were too ambiguous to be added to the lexicon at this point. The remaining unmatched terms were limb pain, renal colic, pain in abdomen, multisite chronic pain, lower limb pain, episodic abdominal cramps, chronic widespread pain, and abdominal cramps.

However, all the pain-related terms (such as cramp, colic, ache, etc.) did match with our lexicon, ensuring the synonyms of pain were indeed all captured.

Medical Subject Headings consist of 63 pain-related MeSH terms and 941 pain-related entry terms. Upon comparison with our lexicon, an overlap of 56 terms (89%) was found with the MeSH terms and 649 terms (69%) with the entry terms. The MeSH terms that did not match (11% i.e., seven terms) were not explicitly related to pain, and included terms such as agnosia [a sensory disorder where a person is unable to process sensory information (Kumar & Wroten, 2022)], pramoxine (a topical anaesthetic), and generic somatosensory disorders. The entry terms that did not match (33% i.e., 307 terms) consisted of drug names (2% of total terms, 5% of non-matched terms) such as Pramocaine and Balsabit, disorders and syndromes (20% of total terms, 62% of non-matched terms) such as visual disorientation syndrome and Patellofemoral syndrome, generic terms (10% of total terms, 31% of non-matched terms) such as physical suffering, and tests (1% of total terms, 3% of non-matched terms) such as Formalin test. The pain specific terms within this list were mainly pain (50% of total terms), -algia (8%), ache (7%), -dynia (1%), and -algesia (1%). Two new pain terms discovered within this list were “catch” and “twinge” which might reference pain in the right context but could also lead to false positives when used in NLP tasks to identify mentions of pain.

3.6 Discussion and Conclusion

When looking at how pain was mentioned in the different text sources, most mentions fell into similar themes, i.e., quality of pain, feelings/experiences associated with the pain, medications, and other measures for pain relief, and mentions of different body parts associated with the pain. The mentions within MIMIC-III were geared more toward pain relief,

which is likely due to the data being from critical care units. In contrast, CRIS covered the feelings and experiences associated with pain. It was hard to get a good sense of the Twitter mentions owing to the short length of strings, while Reddit was a lot more detailed around patient experiences, and pain relief remedies.

The information gained from this exploration helped decide the sources for the development of the pain lexicon. Embedding models built using MIMIC-II/III and CRIS databases were used. The final lexicon consisted of 382 pain-related terms. Embedding models built using Twitter (Pennington et al., 2014) and Reddit (A. Agarwal, 2015) (*Reddit Word Embeddings*, n.d.) data were excluded from inclusion into the final lexicon due to the terms not being very relevant to the term “pain.” They generated terms such as brain, anger, patience, and habit with Twitter, and words such as apartment, principal, and goal by Reddit. The Venn diagrams demonstrated the benefits of including different sources as each of these sources provided unique terms thereby enriching the lexicon for pain. CRIS and MIMIC contributed 68 unique terms that are used in “real-life” settings to the final lexicon. These mostly consisted of commonly used words like soreness, pain, aches. Many of these mentions are potentially based on what patients have said, which could also explain why they are a smaller number of terms. The literature and ontologies have a greater variety of words, as they either use more technical terms, or enumerate every term and concept associated with pain. Apart from helping build the lexicon, this exploration will also help further planning for development of NLP applications and deciding on what attributes around pain might be of interest for general and clinical research purposes.

The final lexicon has been validated by two clinicians, compared to an existing Experimental Factor Ontology which consisted of 26 pain-related terms, and MeSH headings and terms (63 pain-related heading terms and 491 pain-related entry terms). The majority of the pain-related

terms from both these sources matched those included within the lexicon. The terms that did not match were names of disorders/syndromes that may have pain as a symptom, and other more generic words that could lead to false positives if used in downstream NLP tasks.

This study has several limitations. Most importantly, only a small sample of documents was reviewed for the exploration step. Reviewing a larger sample might have been more representative of the text sources and might have revealed deeper insights. The process of exploration of pain concepts within different sources also highlighted the ambiguous nature of a word like pain, and the different contexts that could contain these mentions (metaphorical or clinical mentions). These factors are important to bear in mind when attempting to use such ambiguous terms in NLP tasks as they could lead to false positive results.

The final lexicon, and the code used to generate the embedding models, have been made openly available. This final lexicon will be used in downstream tasks such as building an NLP application to extract mentions of pain from clinical notes which will in turn help answer important research questions around pain and mental health. The findings from this work highlighted that while social media is a vast new data source, it is not always useful in such instances. Social media presents the perspectives of patients which can be beneficial in several ways, however, when constructing a lexicon for use on clinical text, it is essential for the data sources to include language that might be used in a clinical setting. Regardless, this work presents a framework for construction of such a lexicon for clinical terms. The approach followed for the development of this lexicon could be replicated for creating other domain-specific medical lexicons. Future work included patient engagement in order to elicit feedback on the terms that have been included in the lexicon. This lexicon has now been used in follow-up research where it was utilised to help in identification of documents within the CRIS database that mentioned any pain terms. These documents were used to train a machine

learning based application, the details of which are in the upcoming chapters. In addition to this, the lexicon will be formalised for submission to portals, such as BioPortal, for wider use by the community.

3.7 Data Availability Statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/jayachaturvedi/pain_lexicon

3.8 Author Contributions

The idea was conceived by JC, AR, and SV. JC conducted the data analysis and drafted the manuscript. AR and SV provided guidance in the design and interpretation of results. AM provided scripts and guidance on building some of the embedding models. All authors commented on drafts of the manuscript and approved the final version.

3.9 Funding

AR was funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. AR receives salary support from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. JC was supported by the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. AM was funded by Takeda California, Inc. 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 and King's College London. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study received funding from Health Data Research UK, KCL funded CDT in Data-Driven Health, and Takeda California, Inc. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

3.10 Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support. An application for access to the Clinical Record Interactive Search (CRIS) database for this project was submitted and approved by the CRIS Oversight Committee (Oxford C Research Ethics Committee, reference 18/SC/0372). The authors are also grateful to the two clinicians, Dr Robert Stewart and Dr Brendon Stubbs for taking the time to review the terms within the lexicon and providing valuable feedback, as well as Dr Natalia Viani for providing access to some of her python scripts and embedding models.

3.11 Patient and Public Involvement

Patient and public involvement in research is an active collaboration between researchers and members of the public, where the latter actively participate in contributing to the research. This could be as advisers and possibly co-researchers (Research Design Service South Central, 2023).

As part of the development of this lexicon of pain terms, a PPI group of people with lived experiences of SMI and pain, were consulted. The group was asked to review the terms and advise on whether they thought any other written form of communicating pain was missing from the lexicon. For ease of review, the terms were laid out in mini-tables, each consisting of about 5-8 pain terms, grouped by category, such as anatomy-based terms, pain quality-based terms, and so on. The group was given 2 weeks to review the terms, followed by a virtual discussion on Microsoft Teams. They agreed that the lexicon captured terms that they would use to describe their pain. They suggested the addition of one term – “head exploding pain” – which has now been incorporated into the lexicon.

The group shared their thoughts on the importance of this work and suggested research questions that could be answered using such data. Some of these research questions have been incorporated into the prevalence study that was conducted at the end of this project (prevalence study described in Chapter 9) and details of the outputs from the PPI meeting in Section 10.5.3 of Chapter 10.

CHAPTER 4: Sample Size Calculation and Data Extraction

4.1 Foreword

The previous chapter detailed the development of a lexicon of pain terms aimed at facilitating the extraction of pain-related text, for use in downstream classification tasks. This chapter describes work undertaken to determine the sample size for the data required to train the classification models that will undertake the classification tasks.

This work was submitted to two journals (Journal of American Medical Informatics Association and Journal of Biomedical Informatics) and was rejected at both instances. However, valuable feedback was received from the reviewers to improve the robustness of this work. The work presented here, therefore, requires numerous improvements. These improvements will be undertaken as part of future work, which will not be a part of this thesis, and have been detailed in Section 4.12 at the end of the chapter.

4.2 Sample Size Calculation

This section of the chapter describes a simulation of an openly available dataset, with the objective of providing guidelines and recommendations on training data sample sizes and class proportions when building binary classifier models on healthcare data.

The work was conducted in collaboration with Daniel Stahl, Diana Shamsutdinova and Felix Zimmer. The work was presented as proposed work at the Health Text Analytics conference

2022 (HealTAC 2022) and with results at the Health Text Analytics conference 2023 (HealTAC 2023). A link to the preprint is provided below.

Chaturvedi, J., Shamsutdinova, D., Zimmer, F., Velupillai, S., Stahl, D., Stewart, R., & Roberts, A. (2023). ***Sample size in natural language processing within healthcare research*****. https://doi.org/10.2139/ssrn.4553964**

The idea for this work was conceived by me with input from Angus Roberts, Robert Stewart, Daniel Stahl, and Diana Shamsutdinova. Felix Zimmer provided comments on the draft, and suggestions to improve the work. I contributed as the first author, extracted the data, conducted the simulations, and drafted the manuscript.

The following sub-sections of this chapter reproduce the pre-print of the paper, with some minor formatting adjustments to keep it in line with the thesis format. The content itself has not been altered.

Sample Size in Natural Language Processing within Healthcare Research

Jaya Chaturvedi[1*], Diana Shamsutdinova[1], Felix Zimmer[1,3], Sumithra Velupillai[1], Daniel Stahl[1], Robert Stewart[1,2], Angus Roberts[1]

[1]Institute of Psychiatry, Psychology and Neurosciences, King's College London, London, United Kingdom
[2]South London and Maudsley NHS Foundation Trust, London, United Kingdom
[3]Psychological Institute of the University of Zurich, Zurich, Switzerland
*Corresponding author

4.3 Abstract

Objective

Sample size calculation is an essential step in most data-based disciplines. Large enough samples ensure the representativeness of the population and determine the precision of estimates. This is true for most quantitative studies, including those that employ machine learning methods, such as natural language processing, where free text is used to generate predictions and classify instances of text. Within the healthcare domain, the lack of sufficient corpora of previously collected data can be a limiting factor when determining sample sizes for new studies. This paper tries to address the issue by making recommendations on sample sizes for text classification tasks in the healthcare domain.

Materials and Methods

Models trained on the MIMIC-III database of critical care records from Beth Israel Deaconess Medical Centre were used to classify documents as having or not having Unspecified Essential Hypertension, the most common diagnosis code in the database. Simulations were performed using various classifiers on different sample sizes and class proportions. This was

repeated for a comparatively less common primary diagnosis code within the database of diabetes mellitus without mention of complications.

Results

A table listing the expected performances for different classifiers under varying conditions of sample size and class proportion is presented. Smaller sample sizes resulted in better results when using a K-nearest neighbours' classifier, whereas larger sample sizes provided better results with support vector machines and BERT models. Overall, a sample size larger than 1000 was sufficient to provide decent performance metrics (F1 score of 0.80).

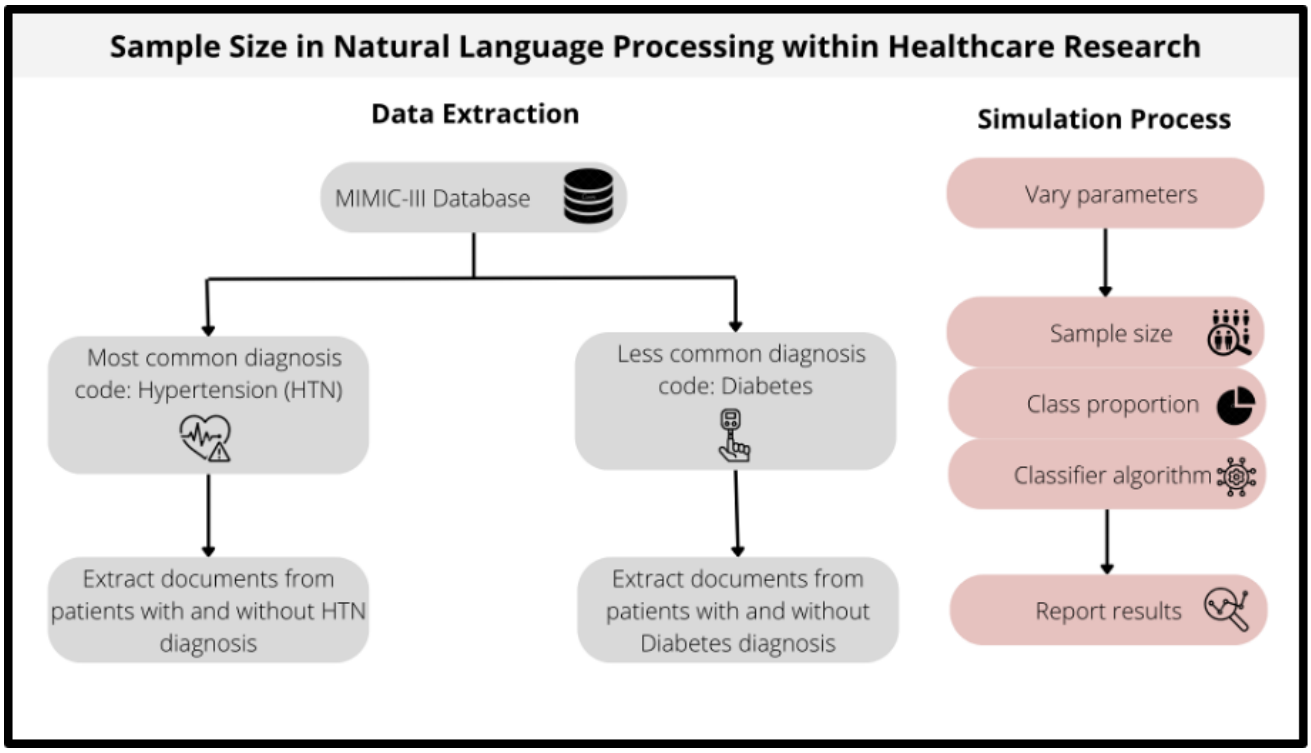
Conclusion

The simulations conducted within this study provide guidelines that can be used as recommendations for selecting appropriate sample sizes and class proportions, and for predicting expected performance when building classifiers for textual healthcare data. The methodology used here can be modified for sample size estimate calculations with other datasets.

Keywords

sample size, natural language processing, machine learning

Graphical Abstract



4.4 Introduction

A sample is a subset of a larger population. The aim of a good sample is for it to be representative of the larger population and provide results approximating what might be found when using the entire population (Ehrenberg, 2000). Knowing whether an appropriate sample size was used is crucial to determine the value of a research project as it gives an insight into whether appropriate considerations were made to ensure the project is ethical and methodologically sound (Faber & Fonseca, 2014).

As with other types of research, an appropriate sample size is essential for quantitative research, especially for the generalisability and reproducibility of the findings (DELİCE, 2010). Most epidemiological studies focus on relationships between some exposure variables and disease outcomes (Cai & Zeng, 2004). In these instances, it is of importance to use a sample that is truly representative of the population of interest, in order to prevent inaccurate deductions from any statistical analyses conducted on these samples. While there are multiple factors that can result in unrepresentative samples, insufficient sample size is a particularly dangerous one. Although random samples are on average unbiased, a small sample will often fail to accurately represent the underlying population, leading to inaccurate observations regarding the relationship between predictors and outcomes. Along with sample size, it is important to use appropriate methods for ascertaining such samples, such as random sampling, in order to avoid magnifying unrepresentativeness within large samples (Msaouel, 2022). Despite this being an important step in quantitative research, a study found that 60% of publications on such research do not provide details on sampling methods and approaches, and if they do, then it is very brief and not reproducible (DELİCE, 2010).

Different research methods demand their own niche sample size calculations. Research methods that utilise natural language processing (NLP) are no exception. NLP is a branch of artificial intelligence within computer science that combines computational linguistics with statistical and machine learning models, enabling the analysis and processing of text data (I B M, 2021). NLP methods are widely used within healthcare research due to the growing volume of data available from electronic health record (EHR) databases (Esteva et al., 2019). EHRs within a single hospital could generate about 150,000 pieces of data (Esteva et al., 2019), some of which may be in textual form. Text is often used for ease of capturing fine-grained details or supplemental information which do not fit into any predetermined structured fields, such as patients' medical histories, preliminary diagnoses, medications, and so on (Ehrenstein et al., 2019).

A prerequisite for a good sample size is the availability of sufficient data to represent the larger population and provide adequate precision in output. This can be quite challenging in the healthcare domain due to the scarcity of openly available healthcare datasets, and the privacy regulations surrounding the use of such data from hospitals and primary care. Data scarcity is compounded by the fact that NLP-based supervised machine learning models require large numbers of human-annotated (in the healthcare domain, ideally clinician-annotated) data as a prerequisite, it can be a limitation due to time and cost constraints (Negida et al., 2019). Insufficient sample sizes within NLP can lead to algorithms that do not perform adequately. We reviewed 11 papers describing past i2b2/n2c2 challenges (spanning from 2006 to 2018) (Pradhan et al., 2015; A. Stubbs et al., 2019; W. Sun et al., 2013; Ö. Uzuner & Stubbs, 2015; O. Uzuner et al., 2007, 2008; O. Uzuner, 2009; O. Uzuner et al., 2010, 2012; Ö. Uzuner et al., 2011, 2017, 2020) widely considered benchmarks in the field of clinical NLP. We found that while sample sizes were described for all training and test sets, a wide range of sample sizes were used (from 288 to 1243), and no justification was provided as to why any of the

sample sizes were chosen. This has been summarised in Table 4.1. We do not say this as criticism specific to these highly cited and regarded papers, but to illustrate that justifications of sample size are rarely given, even in the best clinical NLP studies. This further highlights the need for recommendations on sample sizes specific to this field, and the need for guidelines on what sample sizes are needed, given limited data, to build machine learning models that perform well.

Year	Challenge	Paper	Sample size mentioned	Justification provided?
2006	Deidentification & Smoking	Identifying Patient Smoking Status from Medical Discharge Records (O. Uzuner et al., 2008)	928	No
2007	Deidentification & Smoking	Evaluating the State-of-the-Art in Automatic De-identification (O. Uzuner et al., 2007)	889	No
2008	Obesity	Recognizing Obesity and Comorbidities in Sparse Data (O. Uzuner, 2009)	1237	No
2009	Medication	Extracting medication information from clinical text (O. Uzuner et al., 2010)	1243	No
2010	Relations	2010 i2b2/VA challenge on concepts, assertions, and	826	No

		relations in clinical text (Ö. Uzuner et al., 2011)		
2011	Coreference	Evaluating the state of the art in coreference resolution for electronic medical records (O. Uzuner et al., 2012)	978	No
2012	Temporal Relations	Evaluating temporal relations in clinical text: 2012 i2b2 Challenge (W. Sun et al., 2013)	310	No
2014	Deidentification & Heart Disease	Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks (Ö. Uzuner & Stubbs, 2015)	600	No
2016	RdoC for Psychiatry	A natural language processing challenge for clinical records: Research Domains Criteria (RdoC) for psychiatry (Ö. Uzuner et al., 2017)	1000	No
2018	ADE & Medication Extraction	Advancing the state of the art in automatic extraction of adverse drug events from narratives (Ö. Uzuner et al., 2020)	505	No

2018	Clinical Trial Cohort Selection	New approaches to cohort selection (A. Stubbs & Uzuner, 2019)	288	No

Table 4.1. i2b2/n2c2 challenges – whether justification was provided for the sample sizes used

This table lists out the various i2b2/n2c2 challenges, the tasks involved in the challenges, the sample sizes used in them, and whether any justification was provided within the tasks for the use of these sample sizes.

When referring to machine learning models in general and within NLP, sample size calculations can be used at different stages, such as for training, validation, and testing. Previous work has been conducted to determine sample sizes for validation (Negida et al., 2019; Riley et al., 2020) and limited research has been published on general sample sizes for NLP (Sordo & Zeng, 2005). Sordo et al. (2005) examine the effect of sample size on the accuracy of classification with three classification methods (I Bayes, Decision Trees, and Support Vector Machines) using narrative reports from a hospital, classifying the smoking status of patients (Sordo & Zeng, 2005). They conclude that there is indeed a correlation between the size of the training set and the classification rate, and models show improved performance when trained with bigger samples (Sordo & Zeng, 2005). Using EHR databases, a recent study by Liu et al. (2021) highlighted the Importance of selecting a sample that is unbiased and truly representative of the population to ensure high-quality research (Liu et al., 2021). While the work reported here does not address issues of bias due to the methodologies used to select samples, it does aim to extend previous research and explore the impact of sample sizes and class proportions for a binary classification task.

A substantial proportion of clinical decision-making is dependent on risk-prediction models for health outcomes, which is why the margin of error for such models should be very low and

the performance of such models should be thoroughly validated (Pavlou et al., 2021). Pavlou et al. (2021) investigate the sample size requirements for such validation studies on prediction models (Pavlou et al., 2021). A number of suggestions have been made on the appropriate sample sizes for such validation studies. These include at least 100 events in the validation data (Harrell et al., 1996), or at least 100 events and 100 non-events (Vergouwe et al., 2005). However, such rules of thumb are problematic and do not take into account the model or validation setting (Riley et al., 2020). In response to this, Riley et al. (2020) suggest sample size calculations that incorporate other measures of model performance (such as expected c-statistics and calibration slope) which allows for more tailored sample size calculations based on the models of interest (Riley et al., 2020). The aims for planning a sample size can vary, such as whether the sample size will be sufficient to reach a particular performance metric or to detect differences in performance measures and pre-specified values, or both. This simulation focuses on the former, focusing on the performance metrics for each sample size and class proportion variation, while Pavlov et al. (2021) focused on the latter (Pavlou et al., 2021).

While methods such as power analysis, which are based on strong assumptions, are frequently used in statistical studies (such as prediction modelling) to determine appropriate sample sizes, with the general intention being the larger the sample size the more power associated with the study (Fitzner & Heckinger, 2010), this approach is not transferable to NLP approaches and has therefore been underutilised within NLP (Card et al., 2020). This could be because of the nature of NLP data which does not conform to the standard experiment designs that are used in other studies (Card et al., 2020; Kraemer & Blasey, 2016; Westfall et al., 2014), as also shown in complex statistical modelling (Landau & Stahl, 2013). Determining the sample size in NLP applications is also complicated by the common use of pre-trained models such as BERT (Devlin et al., 2018), which convert text into a

numerical representation, vectorising words in an embedding. These pre-trained models are based on large text corpora such as Wikipedia, or texts with specialised vocabularies, and are readily available in NLP software packages. Using embeddings in a local NLP project translates semantic knowledge of a large corpus to the local documents being classified (Ghannay et al., 2016), which can alter the required sample size or the choice of an optimal classifier. The recommendations made in this paper aim to complement such measures by providing some form of a standard that can be followed when building NLP models.

The performance Indicators commonly used to evaluate and compare different NLP classification algorithms are AUC-ROC, precision, recall, accuracy, and F1-scores (Grandini et al., 2020). AUC-ROC is independent of specific thresholds or cut-offs (Tafvizi et al., 2022), whereas metrics like precision, recall and accuracy are not. Therefore, F1-score and AUC-ROC will be the metrics to compare results from the different simulated classification models in this study.

The aim of this paper is to provide guidelines and suggestions on appropriate sample sizes for training NLP-based machine-learned classification models. This was achieved by conducting straightforward simulations on an openly available healthcare dataset, utilising open-source software and widely used libraries, and building classifiers using varying sample sizes and class proportions as training data. Performances of the different models are compared and recommendations on sample sizes are made based on this. The purpose is not to compare the different classifiers but to recommend sample sizes based on their performances. Despite the simulations being straightforward, they yield valuable information. Some recommendations do exist within the literature for other broader categories of research such as qualitative and quantitative research (Advance HE, 2003; Cai & Zeng, 2004; DELİCE, 2010; Vasileiou et al., 2018), and sample size for validation of machine learning

models (Negida et al., 2019; Riley et al., 2020), predicting sample sizes using learning curve fitting (Beleites et al., 2013; Figueroa et al., 2012; Richter & Khoshgoftaar, 2019), and determining sample sizes in natural language understanding (Chang et al., 2023). In particular, work conducted by Figueroa et al. (2012) uses clinical data for prediction of text classification performances, and Juckett (2012) use a corpus of dictated letters from a pain clinic to determine number of documents required for a gold standard corpus. The approach described here adds to such resources by using simulations on openly available data for recommending sample sizes specific to NLP application development in the healthcare domain. This is particularly important because NLP in the healthcare domain is often conducted on small datasets. The simulations may easily be extended for other parameters and use cases, to generate further recommendations. While these simulations have been conducted on a hypertension and diabetes diagnosis, the diagnosis is not what is being researched and has been used purely as an example. The objective is to understand how sample sizes and class proportions affect performance. We are not trying to generate any new knowledge on the diagnosis used.

4.5 Methods

A series of simulations were conducted on free-text healthcare data from the MIMIC-III database (Johnson et al., 2016). In the simulations, we varied different features of the NLP process, such as the amount of training data, type of classifier algorithm, and prevalence of each class.

4.5.1 Data Source

Medical Information Mart for Intensive Care (MIMIC-III) is an EHR database which was developed by the Massachusetts Institute of Technology (MIT), and made available for

researchers under a specified governance model (Johnson et al., 2016). MIMIC-III contains data on over 58,000 hospital admissions for over 45,000 patients, including about 1.2 million de-identified clinical notes, such as nursing and physician notes, discharge summaries, and ECG/radiology reports (Johnson et al., 2016).

MIMIC-III was chosen for this study due to ease of access, thereby making the study reproducible. MIMIC-III is commonly used in healthcare research (Abhyankar et al., 2014; Lehman et al., 2012; Mayaud et al., 2013; S Velupillai et al., 2015).

4.5.2 Ethics and Data Access

Access to the MIMIC-III database require that the data be handled with care and respect as it contains detailed information about the clinical care of patients. Access was formally requested and granted through the processes documented on the MIMIC-III website[9]. A course protecting human research participants, including HIPAA (Health Insurance Portability and Accountability Act) requirements was completed. A data use agreement outlining appropriate data usage and security standards was submitted.

4.5.3 Data Selection

For the simulations, we designed a simple binary classification task to classify documents within a subset of the database as coming from patients with a particular diagnosis or not. The most common diagnosis code within the database was identified as Unspecified Essential Hypertension, NOS (HTN, ICD-9 code 4019), and made up for 37% of patients within the MIMIC database. The reason for choosing a diagnosis that was so common in the database was so we would have enough data on the cases and non-cases for that diagnosis

[9] https://physionet.org/content/mimiciii/1.4/

in order to run the simulation. However, to test the transferability of this approach, a less common diagnosis code was also used, that of diabetes mellitus without mention of complication (diabetes, ICD-9 code 25000), which made up for 15% of patients within the database. These codes were extracted from the “icd9_code” diagnosis column within the “d_icd9_diagnoses’’ table. None of the other tables within the database contain any diagnosis information. No distinction is made between primary and secondary diagnosis since this information is not available. While the admissions table mentions the diagnosis, this is not coded, i.e., it is mentioned within the free-text and not considered to be the final diagnosis.

The simulated task would therefore be to classify documents as coming from patients diagnosed with HTN or not, and diabetes or not.

A SQL query was run to extract diagnosis, demographics and documents from patients who had the diagnosis of ICD-9 code 4019 (i.e., HTN), along with another subset of patients who did not have the diagnosis of 4019. The initial extraction consisted of 20,000 records from each subset. This is because there were about 20,000 patients with the diagnosis of 4019, and to match this, the same number of patients without this diagnosis code were randomly extracted. A random sample of 5000 was selected for each class. The demographic and document distributions were compared for both classes to ensure they were similar to each other. We assume that this similarity, and the fact that they come from the same dataset, would ensure minimal noise in the data. This process was repeated for patients with and without diagnosis of ICD-9 code 25000 (i.e., diabetes) (Figure 4.1).

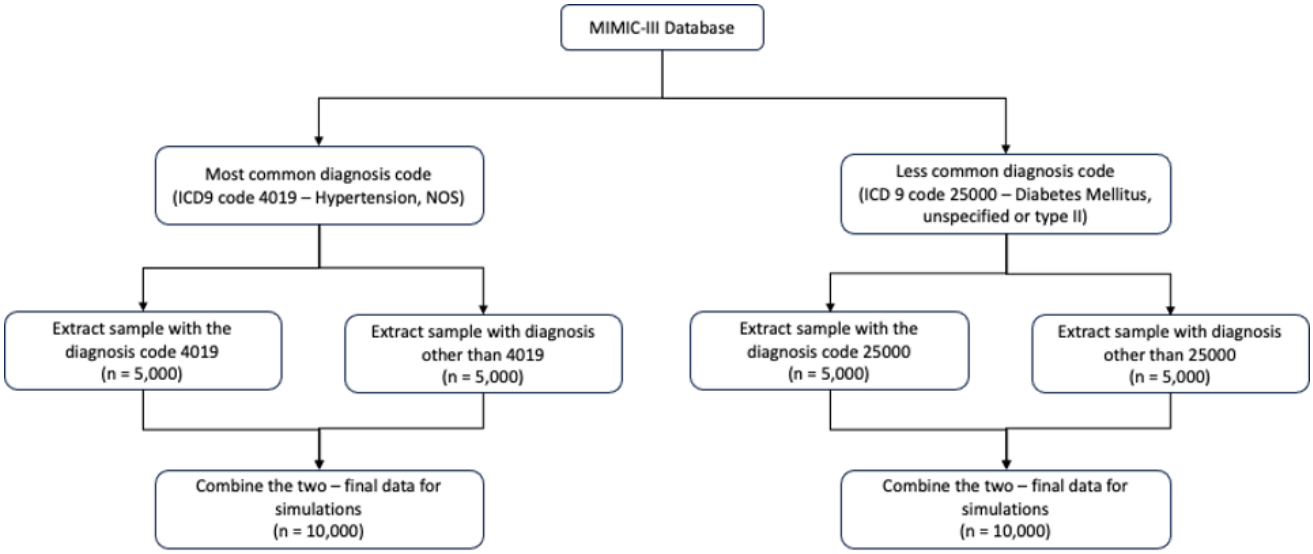


Figure 4.1. Extraction Plan

This flow diagram shows the data extraction plan, where two sets of data were extracted from MIMIC-III, one with the most common diagnosis code within the database i.e., Hypertension, NOS and another with a less common diagnosis code i.e., Diabetes Mellitus, unspecified or type II. These are further split into equal proportions for presence and absence of the diagnosis, and eventually combined to form two datasets for the simulations.

The majority of the notes within the MIMIC-III database are contained within the 'note-events' table (clinical notes such as nursing and physician notes, ECG reports, radiology reports, and discharge summaries) (Johnson et al., 2016). Clinical notes were extracted for both subsets of patients, and the distribution of document lengths calculated and compared.

A typical document within the combined dataset contains information such as dates (admission date, discharge date), history of present illness (explanation of how the patient's current illness developed and any treatment already undertaken), findings from physical examinations, reporting on various tests (scans, blood tests), any diagnosis and treatment plans.

4.5.4 Classifier

For the binary classification task, the diagnosis codes are considered to be the two classes, where patients with a diagnosis code of “4019” are categorised into class 1, and others without this diagnosis code as “class 0”. Similarly, for diabetes, patients with a diagnosis code of “25000” are categorised into class 1, and others without this diagnosis code as “class 0”. This approach for the simulation has been chosen in order to avoid the time-consuming task of manual annotation. The diagnosis codes for patients categorised as class 0 were examined to ensure this group did not contain any diagnosis similar to hypertension or diabetes (a list of all diagnosis codes for class 0 is given in Appendix 4). Thirteen documents within the class 0 of the HTN subset contained hypertension-related diagnosis, such as surgical complication-hypertension (3 patients), malignant hypertension (7 patients), primary pulmonary hypertension (2 patients), and portal hypertension (1 patient). These were removed from the dataset to give the final dataset that was used to run the simulations with varying features as outlined in Table 4.2.

Before building the various classifier models, some pre-processing of the text data was undertaken using the Python NLTK (Natural Language Toolkit) library (Bird et al., 2009). All text was converted to lowercase, any trailing whitespaces, markups such as HTML tags and punctuations/symbols were removed. Common stop words were removed. Words were lemmatised and tokenised. The Python library sklearn (Pedregosa et al., 2011) was used for the non-BERT classifiers, and the Python Huggingface transformers library (version 4.5.0) (Wolf et al., 2019) for the BERT model. The pre-trained BERT_base model was fine-tuned on the simulation data. No steps have been undertaken to balance the data when the class proportions are imbalanced (such as 99/1, 95/5, 90/10, 80/20 class proportions) as the aim was to investigate the effect of this imbalance.

After the sampling into different sizes, each sample was split into train/test/validation sets in the proportions of 60/20/20. The class proportions for each set followed the same prevalence as the main sample size. For example, a sample size of 600 with a 50/50 split would contain 300 documents each in both classes. This was replicated within the train, test and validation sets so the classes were evenly distributed. Distribution of words within each class in the sets were also compared to the overall sample to ensure a similar distribution was represented throughout. For the BERT model, training and validation loss were measured to ensure no overfitting of the models.

Variables	Examples
Size of sample	5000, 4000, 3000, 2000, 1000, 800, 600, 500, 400, 200
Type of classifier/algorithm	Logistic Regression (LR) (Sammut & Webb, 2010), Naive Bayes (NB) (Webb et al., 2010), Random Forest (RF) (Hastie et al., 2009), Decision Tree (DT) (Hastie et al., 2009), Support Vector Classifier (SVC) (Hastie et al., 2001a), Linear Support Vector Classifier (LSVC) (Hastie et al., 2001a), Stochastic Gradient Descent (SGD) (Shai, 2014), K-Nearest Neighbour (KNN) (Hastie et al., 2001b), BERT (BERT_base) (Devlin et al., 2018)
Prevalence of each label	99/1, 95/5, 90/10, 80/20, 70/30, 60/40, 50/50

Table 4.2. Features to be varied

This table lists out the different features that will be varied during the simulations – sample size, classifier, and prevalence of the labels.

The weighted average of F1-score (and confidence intervals) and AUC score were calculated upon varying different features within this classification task. Sample size recommendations have been made, based on the combination of features that perform best.

The queries and code have been made available on GitHub[10].

4.6 Results

4.6.1 Data Summary

MIMIC-III contains 46,520 patients and 58,976 admission records. HTN (ICD-9 code 4019) makes up for 37% (17,613) of all patients and 53% (24,719) of all admissions within the database. Diabetes (ICD-9 code 25000) makes up for 15% (7,370) of all patients and 24% (11,183) of all admissions within the database.

The demographic and document distributions were compared to ensure they were similar enough to each across both classes in the respective subset groups i.e., HTN and diabetes. The demographic distributions are outlined in Table 4.3.

Demographic	HTN cohort		Diabetes cohort	
	Class 0	Class 1	Class 0	Class 1
Gender	Male: 55%	Male: 59%	Male: 56%	Male: 62%

[10] https://github.com/jayachaturvedi/sample_size_in_healthcare_NLP

	Female: 45%	Female: 41%	Female: 44%	Female: 38%
Age*	Mean: -61 Min: -169 Max: 221	Mean: -55 Min: -151 Max: 221	Mean: -68 Min: -179 Max: 221	Mean: -58 Min: -138 Max: 220
Ethnicity	White: 66% Unknown: 12% Black: 9% Hispanic/Latino: 4% Multi-race/Other: 4% Asian: 5%	White: 68% Unknown: 14% Black: 9% Hispanic/Latino: 4% Multi-race/Other: 3% Asian: 2%	White: 67% Unknown: 10% Black: 10% Hispanic/Latino: 6% Multi-race/Other: 3% Asian: 4%	White: 60% Unknown: 17% Black: 9% Hispanic/Latino: 7% Multi-race/Other: 3% Asian: 4%

Table 4.3. Demographic distributions for both classes in both cohorts

**Ages within MIMIC-III may be negative. For the purposes of de-identification and keeping in line with the HIPAA regulations, the database providers have shifted any dates within the database (including age) into the future by a random offset for each individual patient. This has been done in a consistent manner so that the intervals between stays and discharge are preserved. However, due to this shift, hospital stays appear to occur between the years 2100 and 2200, leading to calculated ages being negative. For patients with a date of birth over 89, their age within the database appears as being over 100 years old (Johnson et al., 2016).*

Some patients within class 1 contained hundreds of documents, and so have been eliminated by the application of a limit to the number of documents per patient, in order to maintain consistency. This limit was set to less than or equal to 50 documents per patient. After applying the threshold of 50 documents, the document distribution is displayed in Figures 4.2a for HTN and 4.2b for diabetes cohort.

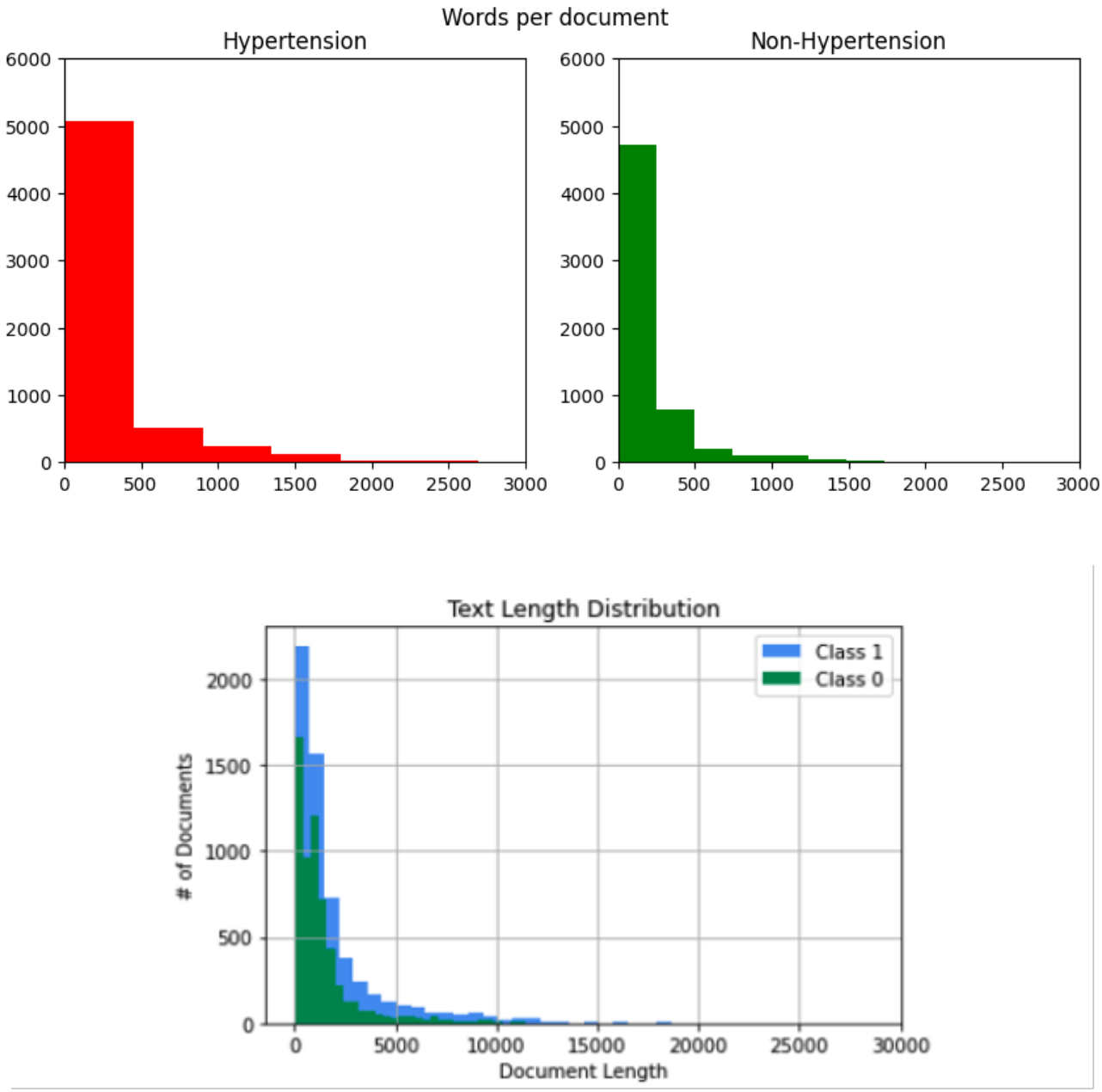
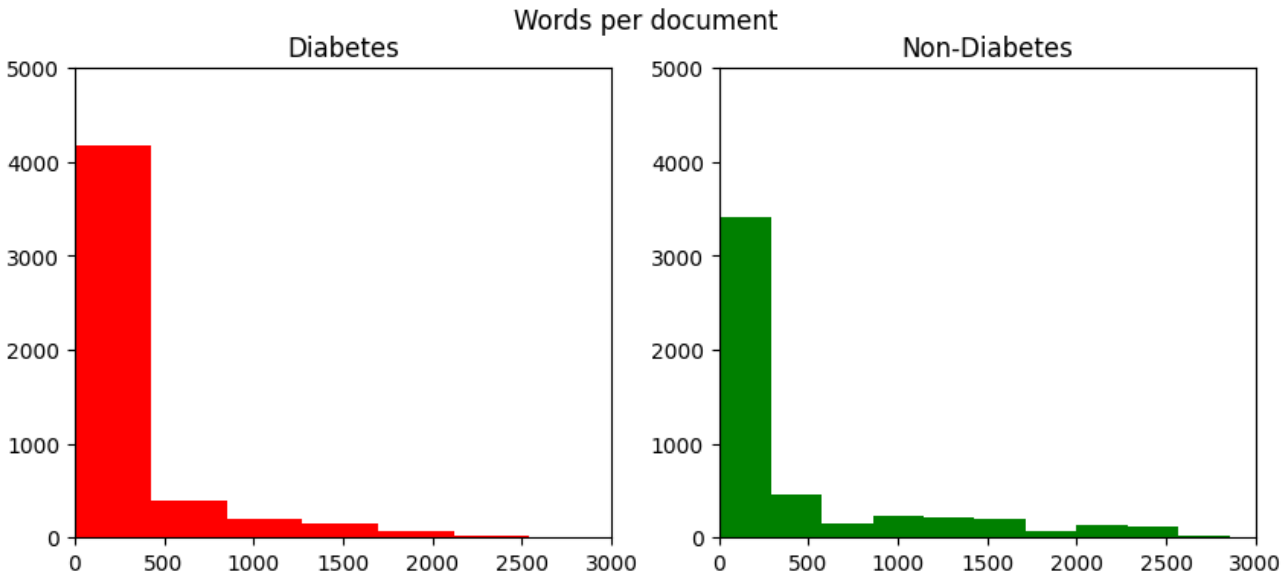


Figure 4.2a. Document distribution between the two classes - Hypertension
These graphs show the words per document (on top) for both the classes i.e., hypertension and non-hypertension, and the document lengths (bottom) for both classes.



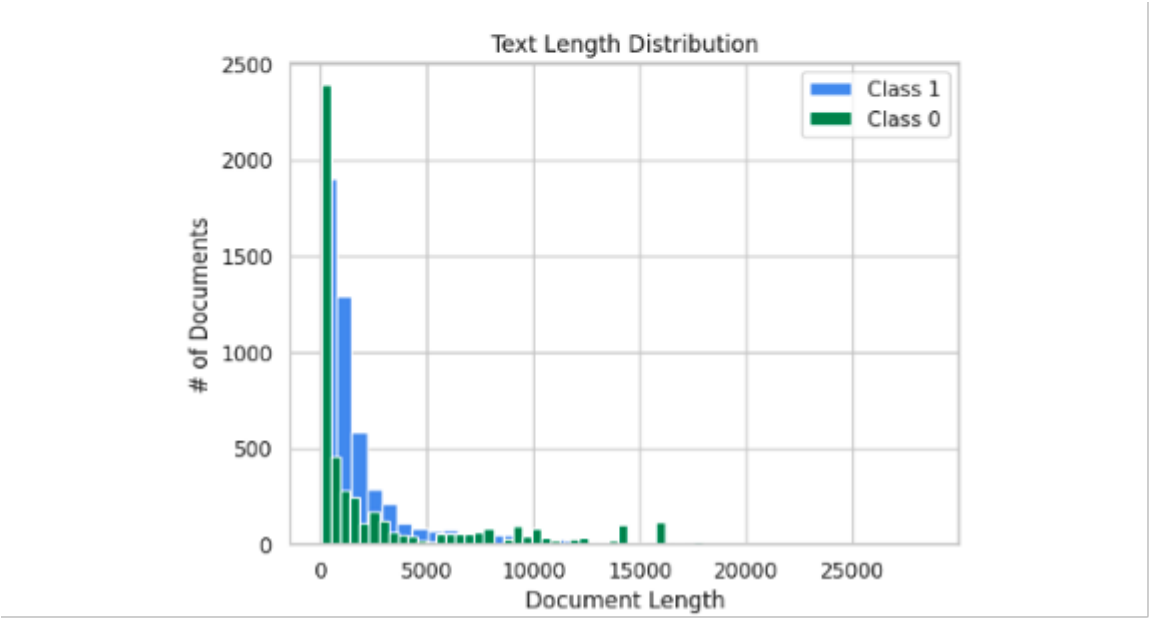
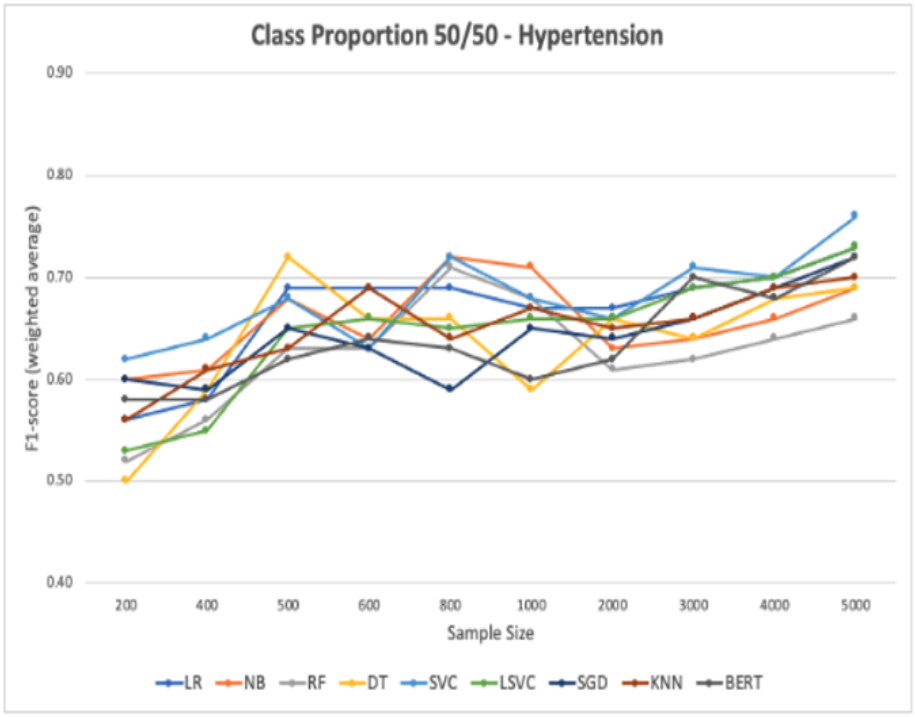


Figure 4.2b. Document distribution between the two classes - Diabetes
These graphs show the words per document (on top) for both the classes i.e., diabetes and non-diabetes, and the document lengths (bottom) for both classes.

4.6.2 Classifier

Effect of variations on Classifier Performance

Figure 4.3 shows the impact of variations in sample size and classifiers on one class proportion variation (50/50)



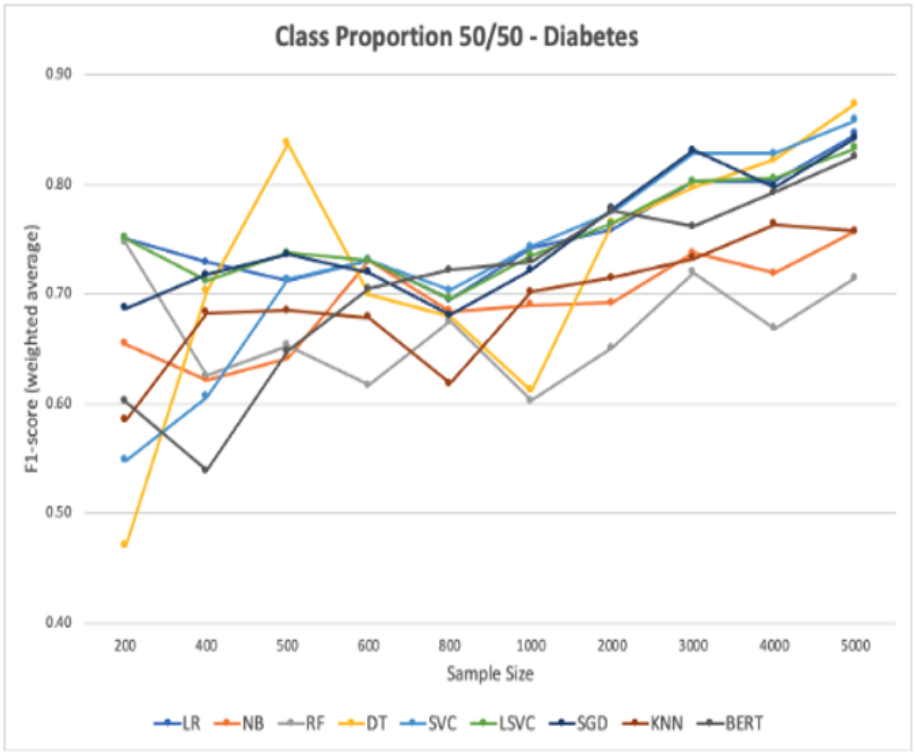


Figure 4.3. F1 for each classifier and different sample sizes at 50/50 class proportion - HTN and diabetes – on the validation set.

A table showing the metrics for all the classifiers and sample size/class proportion variations for both the diagnosis subsets is given in Appendix 2 and Appendix 3, with the range of AUC

and F1 scores for some sample sizes and class proportions of the HTN and Diabetes subsets summarised in Table 4.4. The best performing classifier given in this table was based on the F1 score. AUC scores lead to a different classifier performance ranking. Class proportions of 95/5 and 99/1 have not been included below, but are available in Appendix 2 and Appendix 3.

Range of AUC and F1 scores (Range of AUC [classifier with highest AUC] Range of F1 scores [classifier with highest F1 score])					
Hypertension					
Sample Size	Class Proportion				
	90/10	80/20	70/30	60/40	50/50
5000	0.63 - 0.73 [LR] 0.87 - 0.91 [LSVC]	0.65 - 0.78 [LR] 0.76 - 0.83 [SVC]	0.67 - 0.76 [SVC] 0.67 - 0.78 [LR]	0.67 - 0.82 [SVC] 0.63 - 0.76 [SVC]	0.69 - 0.83 [SVC] 0.66 - 0.76 [SVC]
3000	0.61 - 0.74 [SVC] 0.85 - 0.91 [LSVC]	0.60 - 0.75 [LR] 0.74 - 0.81 [LR]	0.66 - 0.77 [LSVC] 0.66 - 0.78 [LSVC]	0.62 - 0.75 [LSVC] 0.59 - 0.71 [BERT]	0.63 - 0.77 [SVC] 0.61 - 0.71 [SVC]
1000	0.50 - 0.75 [SVC]	0.54 - 0.72 [SGD]	0.53 - 0.63 [RF]	0.60 - 0.71 [SGD]	0.59 - 0.74 [LR]

	0.79 - 0.88 [BERT]	0.74 - 0.83 [SGD]	0.64 - 0.72 [SVC]	0.58 - 0.67 [SGD]	0.59 - 0.71 [NB]
500	0.43 - 0.61 [KNN]	0.59 - 0.78 [LR]	0.59 - 0.71 [NB]	0.64 - 0.73 [SVC]	0.59 - 0.77 [LR]
	0.85 - 0.94 [KNN]	0.67 - 0.86 [LSVC]	0.72 - 0.76 [LR]	0.58 - 0.71 [LR]	0.62 - 0.72 [DT]
200	0.30 - 0.63 [NB]	0.48 - 0.65 [DT]	0.31 - 0.51 [NB]	0.47 - 0.78 [SVC]	0.34 - 0.68 [SGD]
	0.85 - 0.91 [KNN]	0.58 - 0.84 [KNN]	0.55 - 0.73 [SVC]	0.40 - 0.74 [BERT]	0.35 - 0.62 [NB]
Diabetes					
5000	0.68 - 0.87 [LR]	0.72 - 0.88 [LR]	0.79 - 0.91 [SVC]	0.79 - 0.92 [SVC]	0.81 - 0.93 [SVC]
	0.86 - 0.93 [LSVC]	0.75 - 0.88 [SGD]	0.64 - 0.87 [SVC]	0.64 - 0.86 [SVC]	0.71 - 0.87 [DT]
3000	0.58 - 0.88 [LR]	0.70 - 0.85 [LR]	0.74 - 0.88 [LR]	0.76 - 0.90 [SVC]	0.76 - 0.90 [SVC]
	0.87 - 0.93 [LSVC]	0.75 - 0.87 [SGD]	0.67 - 0.84 [SVC]	0.66 - 0.84 [SVC]	0.72 - 0.83 [SGD]
1000	0.50 - 0.84 [LR]	0.50 - 0.80 [LR]	0.73 - 0.87 [RF]	0.70 - 0.84 [LR]	0.61 - 0.85 [LSVC]

	0.83 - 0.89 [SGD]	0.68 - 0.76 [KNN]	0.71 - 0.81 [KNN]	0.59 - 0.78 [SGD]	0.60 - 0.74 [LSVC]
500	0.50 - 0.79 [RF]	0.59 - 0.83 [RF]	0.52 - 0.86 [LSVC]	0.62 - 0.86 [LR]	0.67 - 0.84 [DT]
	0.72 - 0.88 [SGD]	0.75 - 0.87 [KNN]	0.50 - 0.78 [SGD]	0.62 - 0.80 [SGD]	0.64 - 0.84 [DT]
200	0.45 - 0.79 [NB]	0.50 - 0.83 [LR]	0.50 - 0.89 [RF]	0.59 - 0.84 [LR]	0.47 - 0.87 [RF]
	0.80 - 0.85 [BERT]	0.63 - 0.86 [KNN]	0.45 - 0.84 [LSVC]	0.36 - 0.69 [KNN]	0.47 - 0.75 [LSVC]

Table 4.4. Range of AUC and F1 scores for each sample size and class proportion, with the best performing classifier mentioned in brackets.

4.7 Discussion

The simulations conducted in this project are aimed at providing some recommendations on sample sizes that will be useful when building a classifier, based on class proportions and classifier types. As expected, classifiers built with larger sample sizes showed better performance in general, with the recommendation being that a sample size of 1000 or more will generate decent performance (F1 score of 0.80). The results showed that larger sample sizes resulted in some classifiers performing better, and others with more balanced classes. For example, smaller samples (800 and below) resulted in better performance by the KNN classifier and more frequently with imbalanced class proportions such as 90/10, 80/20 and 70/30, when compared to larger samples (1000 and above) that generated better performance with the BERT model. This might be because KNN is a distance-based algorithm

and is able to calculate distances within small datasets with ease (Uddin et al., 2022). KNNs are known to not perform well with large datasets due to the curse of dimensionality where the distance functions that are used within KNNs are rendered ineffective due to high dimensionality within larger datasets (Kouiroukidis & Evangelidis, 2011). Apart from this, the computational costs of calculating distances between new points within the dataset is very high too (Uddin et al., 2022). For some classifiers, such as SVC and LSVC, larger samples are required in order to obtain good performance results. This is interesting because SVCs are generally not considered ideal for large datasets because they are slow to train when using large datasets that also contain large number of features and variations, making it computationally infeasible (Huang et al., 2018). Imbalanced classes performed best most frequently with the BERT model. This could be because the model utilises transfer learning and being pre-trained on a large corpus of language, its architecture produces pre-trained context-dependent embeddings which encode aspects of general language, as shown by the fact that they have proven powerful in solving a multitude of NLP tasks, including handling imbalanced classification tasks without need for any further augmentation or manipulation (Madabushi et al., 2020). Finally, although we compared a number of algorithms, our list is far from being exhaustive, and future work may include a wider range of deep learning and penalised regression models.

As the sample sizes reduce within the simulations, the confidence intervals get wider (reported in Appendix 2 and Appendix 3), as expected. Similarly, the AUC scores are lower with the smaller sample sizes compared to larger ones and often approach on average chance classification (AUC=0.5) i.e., it cannot separate between classes (Hajian-Tilaki, 2013). This is attributed to the accuracy of the classifier rather than the data (sample size and class proportions). Since the accuracy of a classifier is affected by sample size and class proportions, it in turn affects the AUC score.

An important consideration in creation of labelled training data in the real world is the human annotation process, and inter-annotator agreements (IAA). This work does not account for IAA scores, since we are using the diagnosis code as the label, which was potentially assigned by a single coder within the hospital and is being used as the gold standard in our work. An assessment of clinical coding within routinely collected hospital data was conducted by Dixon et al. (1998) which found that inter-coder agreement varied between different medical conditions (Dixon et al., 1998). While this work has focussed on sample sizes, and not addressed the issues of IAA, they are both important factors that might determine the performance of an NLP classifier.

Lastly, the data heterogeneity within healthcare text is a known challenge when it comes to transferability of results and determining whether the recommendations made in this paper can be applied to data from another healthcare database. Since the results obtained in this paper were from data of a critical care unit based in the United States, there will be differences in the structure and content of the textual data when compared to other sources of health data. Even within the same database, the results might vary when a classification task is conducted on another diagnosis code. However, the vocabulary used within different healthcare datasets would have some overlap due to the common terminologies used, so these results can prove to be a useful guide.

While this sample size simulation study provides initial insights, there are certain inherent limitations that warrant the need for further experimentation to derive more robust recommendations. The current methodology and results, though informative, remain inconclusive and would need refinement. Several fundamental aspects of the experimental design and methodology could be enhanced to support the robustness of the findings. It is important to acknowledge these limitations and exercise caution when interpreting the current

results as definitive guidelines for sample size determination. While the work presented here provides a valuable starting point and highlights the importance of this consideration, further experimentation and methodological enhancements are necessary to gain greater confidence in the recommendations. Detailed plans to address these issues and extend the sample size simulation study are outlined in Section 4.9. However, given the scope and complexity of these proposed improvements, they will be pursued as part of future work, beyond the confines of this thesis. Nonetheless, the current findings underscore the significance of this research avenue and lay the groundwork for more comprehensive and robust investigations to follow.

4.8 Conclusion

This paper provides recommendations on sample size for training data when building classification models. It was found that a minimum sample size of 1000 could generate a decent F1 score of 0.80 or above. These recommendations are based on simulations that were conducted on the MIMIC-III dataset, using patient documents with the most common diagnosis code (HTN) as class 1 and a similar cohort of patient documents with any other diagnosis code as class 0. The sample sizes were varied incrementally from 200 to 5000 documents, and class proportions varied from a 50/50 split of classes to a 90/10 split. Different classification algorithms were used on these varying sample sizes and class proportions. This simulation was repeated with a less common diagnosis code (diabetes), as a test of the transferability of this approach. The results have been reported briefly in Table 4.4 and in more detail in Appendix 2 and Appendix 3. The objective is to use these recommendations as guidelines when conducting similar classification tasks within the healthcare domain.

While it is not unusual for reports of clinical NLP to compare results for different parameters and algorithms with respect to specific narrow research questions, to the best of our knowledge this paper is the first to report a reproducible methodology and guidelines using open tools and data, for the purpose of guiding general decisions on sample size across a range of NLP research. The recommendations from this work could be applicable to classification tasks being conducted on other similar datasets within healthcare. One of the limitations of this work is that this was carried out on a critical care unit database which contains specific terminologies and potentially more abbreviations within their notes, which might not be as transferable to a different kind of dataset, such as health records from other medical specialities, which might contain different styles and complexities of narrative. However, the methodology presented here can be replicated by other researchers for sample size estimations using their own datasets. An understanding of appropriate sample size will enable researchers to better judge both the replicability and reproducibility of reported studies and therefore to understand the limitations of those studies.

As mentioned in the previous section, further refinement of the methodology presented here is required and will be undertaken as future work (detailed in Section 4.9). In addition to these refinements, we also intend to conduct experiments where we vary the level of classification (such as sentence or token level vs. document level). This would have to overcome the lack of availability of labelled health record data for some levels, such as tokens. We will also consider the split between train/test/validation sets within a sample, and the type of classification (such as multiclass vs. binary). We plan to run a simulation on a different dataset to test the transferability of our approach. We also plan to expand our simulation approach to investigate if such text features as the size of the underlying vocabulary, number of words per document, or similarity of the descriptive words for the positive and negative classes, affect the minimum training sample size required for an NLP model. The methodology proposed

here has provided guidelines and recommendations on what sample sizes and class proportions should be used for binary health record document classification. The same methodology could be used for future extensions to this work.

4.9 Future Work

As mentioned previously, certain flaws were identified in this work and will require a substantial amount of rework to make the results more robust. These improvements will be undertaken as part of future work, not included as part of this thesis. The main changes are outlined below.

1. Extraction of more data and conducting more simulations – More data will help with conducting more simulations, with more gaps between them, until performance plateaus.
2. Variation of class proportions - In its present state, the simulations use varying proportions for class 1 (such as 50%, 60%, and so on until 99%), but a similar variation has not been performed on class 0. Using larger proportions of the negative class (class 0) is more realistic when studying rare conditions, or diagnosis that is less common, and therefore essential to incorporate in these simulations.
3. The aim and scope of this work can focus on achieving a fixed performance metric, such as F1 score of 0.80, instead of aiming for “decent performance” which is vague.
4. Ensuring balance within the different samples in terms of diversity and representativeness. The current work shows distributions by age, gender and ethnicity, which are not too dissimilar between the classes, but there is scope to balance it even more. Stratified sampling can be applied to ensure such balance.
5. Comparing the performance of these simulations to other work that showcases state of the art performance on other biomedical data.

6. Repeating the simulations for each variation about 10 times and average the performance, showing how wide or narrow the boundaries of the performance can be. This can be achieved by bootstrapping, and could increase the smoothness of the metric graphs which are quite uneven in the current work.
7. Use a wider range of diagnosis codes for hypertension and diabetes respectively, to minimise the chances of related diagnosis codes being included as part of class 0. Diagnosis codes described by Birman-Deych et al. (2005) can be used as guidance to decide on the ICD-9 codes.

4.10 Authors' Contributions

The idea was conceived by JC, AR, DS, RS and DSham. JC conducted the simulations and drafted the manuscript. AR, RS, DS, DSham, SV, and FZ provided guidance in the design and interpretation of results. All authors commented on drafts of the manuscript and approved the final version.

4.11 Funding Statement

AR was part-funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. AR's and RS's salaries were part supported by the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities. DS, RS and AR are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS

Foundation Trust and King's College London. RS is additionally part-funded by the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust, and by the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386). FZ was funded by the Swiss National Science Foundation (SNSF grant 188929). JC was supported by the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. DSh is funded by the King's College London Biostatistics and Health Informatics PhD studentship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.12 Declaration of Competing Interests

RS declares research support received in the last 3 years from Janssen, GSK and Takeda. AR declares research support received in the last 3 years from Takeda and RSM UK. The other authors have no competing interests to declare.

CHAPTER 5: Development of a Corpus Annotated with Mentions of Pain in Mental Health Records

5.1 Foreword

Having described my work on developing a lexicon for pain entities (Chapter 3), and work carried out more generally to estimate sample sizes for gold standard annotations, and the extraction of data from CRIS for the creation of these annotations (Chapter 4), this chapter moves on to describe the development of a corpus annotated with mentions of pain, including details of the annotation process and the description of the annotated gold standard corpus.

The work was conducted in collaboration with three medical student annotators - Natalia Chance, Luwaiza Mirza and Veshalee Vernugopan. I contributed as the first author, designed the research, developed the annotation guidelines in collaboration with the annotators, calculated annotation agreement scores, analysed distributions of the annotations, and drafted the manuscript. This work has been published in the *Journal of Medical Internet Research*.

Chaturvedi J, Chance N, Mirza L, Vernugopan V, Velupillai S, Stewart R, Roberts A. ***Development of a Corpus Annotated With Mentions of Pain in Mental Health Records: Natural Language Processing Approach.*** **JMIR Form Res 2023; 7: e45849. URL: https://formative.jmir.org/2023/1/e45849. DOI: 10.2196/45849**

The following sub-sections of this chapter reproduce the paper, with some minor formatting adjustments to keep it in line with the thesis format. The content itself has not been altered. More details about the disagreements between the annotators during the process have been detailed in Appendix 5.

Since the publication of this manuscript, two related works have been identified (Dave et al., 2022 and Naseri et al., 2023), albeit not in the mental health context, and so have been added as citations at the end of the Introduction section. In addition, more details on the data extraction process have been included.

Development of a Corpus Annotated with Mentions of Pain in Mental Health Records

Jaya Chaturvedi[1]*, Natalia Chance[1], Luwaiza Mirza[1], Veshalee Vernugopan[2], Sumithra Velupillai[1], Robert Stewart[1,3], Angus Roberts[1,3]
[1]King's College London, [2]University of Glasgow, [3]South London and Maudsley Biomedical Research Centre
*corresponding author

5.2 Abstract

Pain is a widespread issue, with 20% of adults suffering globally. A strong association has been demonstrated between pain and mental health conditions, and this association is known to exacerbate disability and impairment. Pain is also known to be strongly related to emotions, which can lead to damaging consequences. As pain is a common reason for people to access healthcare facilities, electronic health records (EHRs) are a potential source of information on this pain. Mental health EHRs could be particularly beneficial since they can show the overlap of pain with mental health. Most mental health EHRs contain the majority of their information within the free-text sections of the records. However, it is challenging to extract information from free text. Natural language processing (NLP) methods are therefore required to extract this information from the text. This research describes the development of a corpus of manually labelled mentions of pain and pain-related concepts from the documents of a mental health EHR database, for use in the development and evaluation of future NLP methods. The EHR database used, CRIS (Clinical Record Interactive Search), consists of anonymised patient records from The South London and Maudsley (SLaM) NHS Foundation Trust in the UK. The corpus was developed through a process of manual annotation where pain mentions were marked as relevant (i.e., referring to physical pain afflicting the patient), negated (i.e., indicating absence of pain) or not-relevant (i.e. referring to pain affecting someone other than

the patient, or metaphorical and hypothetical mentions). Relevant mentions were also annotated with additional attributes such as anatomical location affected by pain, pain character, and pain management measures, if mentioned. Over 70% of the mentions found within the documents were annotated as relevant, and about half of these mentions also included the anatomical location affected by the pain. In future work, the extracted information will be used to develop and evaluate a machine learning based NLP application to automatically extract relevant pain information from EHR databases.

Keywords

Pain, Mental Health, Natural Language Processing, Annotation, Information Extraction

5.3 Introduction

Pain is a growing focus of research, especially since the opioid crisis in the United States (Howard et al., 2018). Pain can have long-term implications on the emotional well-being and mental health of people (Heintzelman et al., 2013) due to its debilitating nature, and has a potential impact on healthcare and societal costs (Groenewald et al., 2014). Pain is known to affect one in five people (International Association for the Study of Pain, 2005), is a common reason for people to access healthcare facilities, and therefore features in patients' health records. For these reasons, a potential source of information on pain is electronic health records (EHRs), which contain rich data on interactions between clinicians and patients (Viani et al., 2021).

Mental health EHRs are particularly beneficial for this research since they have the potential to show the recorded overlap of pain with mental health in clinical encounters. EHRs generally consist of structured information (such as tables and forms) and unstructured information (such as correspondence letters and discharge summaries). In mental health EHRs, the majority of information lies within the unstructured and free-text sections of the records (Viani et al., 2021). This may be because free-text fields allow clinicians the flexibility required to capture pertinent information on patient experiences, which might not be possible in the structured fields, which contain mostly drop-down menus and predetermined options that would not fit every patient situation.

Tian et al. (2013) have presented work in which they attempted to identify patients with chronic pain from EHRs in a primary care setting. They used structured information from the EHR, including a combination of diagnostic codes for potential chronic painful conditions, patient-reported pain scores, and opioid prescription medications to identify patients (Tian et al.,

2013). Their research has highlighted that pain is not captured very well in the coded structured fields of EHRs, thereby making the free-text sections a valuable resource for extracting this information. However, since the description of pain is quite ambiguous in nature, it is challenging to extract accurate information about pain from text. Natural language processing (NLP) methods offer a potential solution, employing computational methods for the analysis of linguistic data, aiding in the efficient extraction of relevant pain information from clinical documents.

The aim of the research described here was to develop a corpus of mentions of pain from the documents of a mental health EHR database called CRIS (Clinical Record Interactive Search), which consists of anonymised patient records from The South London and Maudsley (SLaM) NHS Foundation Trust in the UK, one of the largest mental health care providers in Western Europe (Stewart et al., 2009). Documents containing mentions of pain were identified and manually annotated with a number of different pain attributes, thereby creating a human-labelled dataset. The annotation was conducted using the information immediately surrounding a pain term (200 characters before/after the term) within a document, i.e., the context of the entire document was not considered for the annotation task. Sentences within this task are determined as 200 characters before/after a pain term, and not actual sentences. This was done because it is challenging to split the text into sentences (lack or inconsistent use of punctuations) due to the messy nature of the textual data within EHRs. In future work, sentences from these labelled documents were used to develop NLP applications to automatically extract such information from EHR databases (Chapter 6). Along with development of the annotated corpus, this paper also investigates the distribution of pain and its different attributes within these mental health records. While recent work has been done to extract pain information from EHRs (Dave et al., 2022; Naseri et al., 2023), to the best of

our knowledge, such an extraction of pain information from clinical text of mental health EHRs has not previously been conducted.

5.4 Methods

5.4.1 Data Access Statement

The source clinical data are accessed under the auspices of SLaM, the data custodian. The Maudsley CRIS platform provides access to anonymised data derived from SLaM's electronic medical records within a bespoke information governance framework[11]. These data, and any NLP application built using this data, can only be accessed by permitted individuals from within a secure firewall (i.e., the data cannot be sent elsewhere), under the same controlled conditions for data security and privacy followed by the authors. All access is password protected using authenticated SLAM network usernames and passwords. Data exported from CRIS cannot be saved directly outside the firewall. Further technical details of the security model can be accessed on their website[12].

The CRIS application was approved as a database for secondary analysis by the Oxford Research Ethics Committee (18/SC/0372). A patient-led oversight committee (detailed in (Fernandes et al., 2013)) reviews and approves research projects that use the CRIS database. Service users are actively involved in the development of the CRIS database and manage the strict governance frameworks related to it. All SLaM patients are given the opportunity to opt out of their data being used for purposes other than their care (Fernandes et al., 2013; Ford et al., 2020; NHS Digital, 2022a). Data are owned by a third party, Maudsley

[11] For more information please contact: cris.administrator@slam.nhs.uk
[12] https://www.maudsleybrc.nihr.ac.uk/media/112184/cris-security-model.pdf

Biomedical Research Centre (BRC), which runs the CRIS tool, providing access to anonymised data. The work undertaken as part of this project, as well as related research, were all approved by the CRIS Oversight Committee (CRIS projects: 21-021, 23-003).

5.4.2 Data Source

The CRIS database, described in more detail in Section 2.1 of Chapter 2, was the main data source for this project. The CRIS data platform consists of de-identified records from SLaM, one of the largest mental health care providers in Western Europe (Stewart et al., 2009). It consists of Trust-wide records from 2008 to date and is supported by the NIHR Biomedical Research Centre at SLaM and King's College London (Stewart et al., 2009). CRIS follows a robust, patient-led governance model and has ethical approval for secondary analysis (Oxford C Research Ethics Committee, reference 23/SC/0257). The free-text within CRIS is composed of progress notes, discharge summaries, written assessments, correspondence documents, and more. There are over 30 million case notes within this database, averaging about 90 documents per patient (Velupillai et al., 2018). Documents are stored in one of several CRIS database tables, depending on the purpose and type of the document. For example, tables exist for correspondence documents, and for patient encounter event note documents.

5.4.3 Data Extraction

EHR structured tables and codes do not necessarily include information about pain, potentially due to it being a symptom rather than a diagnosis, making it difficult to extract documents based on codes alone. Therefore, this information was sought from the unstructured free-text fields of the database. In order to determine which tables might contain most of the patient documents, a preliminary extraction of 1,000 documents from each table

was conducted (Table 5.1). This selection of 1,000 documents was random, and not related to the recommendations made in Chapter 4. As seen in the table, some sources have very few characters. Upon exploration, it was found that these documents contained text such as “Invalid-Error”, “Letter to GP” or just “NA”.

Table Name	% of records with text	Average number of documents per patient	Length in Characters			
			Mean	Median	Minimum	Maximum
Attachment	100%	1 (max 15, min 1)	11,637	6,562	32	520,678
POSProforma (Nurse Assessment Notes)	100%	1 (max 10, min 1)	10,194	7,460	547	67,498
POSProforma (SPR Assessment Notes)	100%	1 (max 10, min 1)	5,528	3,789	84	30,413
POSProforma (SHO Assessment Notes)	100%	1 (max 10, min 1)	7,021	3,508	73	25,505
Discharge Notification Summary (Brief Summary)	100%	1 (max 8, min 1)	3,058	35	3	5,768
Discharge Notification Summary (Discharge Plan)	100%	1 (max 8, min 1)	412	22	3	5,768
POSProforma (Discharge Notes)	100%	1 (max 10, min 1)	304	126	62	5,191
Event	100%	17 (max 32, min 1)	641	415	1	132
Discharge Notification Summary (Medication)	100%	3 (max 24, min 1)	392	158	15	2,208
CCS Correspondence	99%	2 (max 17, min 1)	261	-	28	17,116

Risk event	99%	3 (max 32, min 1)	508	84	4	6,335
Presenting Circumstances	98%	1 (max 10, min 1)	1,452	933	6	18,122
Mental State Formulation (Mental State Comments)	96%	1 (max 14, min 1)	1,465	1,074	81	21,370
CAMHS Event	88%	2 (max 23, min 1)	919	523	2	10,356
CCS Correspondence (Attachment Text)	81%	2 (max 13, min 1)	2,171	-	7	32,759
POSProforma (Nurse Assessment Presenting Circumstances)	79%	1 (max 7, min 1)	315	311	4	2,515
Mental State Formulation (Assessments Summary Comments)	68%	1 (max 9, min 1)	662	450	76	8,620
Care Plan Mental Health Table	55%	4 (max 44, min 1)	206	154	2	19,634
Care Plan Physical Health Table	40%	2 (max 12, min 1)	283	32	3	8,134
History (Personal History)	28%	1 (max 5, min 1)	949	719	1	8,452
POSProforma (AMHP Assessment Notes)	28%	1 (max 4, min 1)	587	1,259	2	14,313
History (Drug and Alcohol)	20%	1 (max 4, min 1)	186	87	71	3,112

Discharge Notification Summary (Comments)	16%	1 (max 3, min 1)	262	175	1	36,740
POSProforma (S12 Doctor Assessment Notes)	7%	1 (max 2, min 1)	59	23	4	6,552
POSProforma (Diversion Notes)	7%	1 (max 5, min 1)	69	118	4	3,618

Table 5.1. Summary of textual sources within CRIS

This table summarises the various sources of clinical text within CRIS, providing details on the percentage of records within each source that contain textual data, the average number of documents per patient in each source, and the median, minimum, maximum and mean length of characters within the documents. The median was calculated at a later date, and some tables have been made redundant and removed from the database – due to this, information on the median is not available for some tables.

Most discharge notification summaries and correspondence letters are included within Attachments. These also contain the longest documents among all the tables. The Event table records essential face-to-face interactions between the clinician and the patient and holds the highest average number of documents per patient. For these reasons, the Attachment and Event tables were chosen as the sources for the extraction of documents.

In addition to this, a search was conducted on all the text sources within CRIS for the word “pain” to gauge where information about pain might be recorded most frequently. The majority of mentions of pain were in documents from CRIS “Event” and “Attachments” tables (Table 5.2), and so these were used in the next steps. In SLaM EHRs, “Event” documents represent conventional case notes, usually completed by the reviewing member of staff contemporaneously with, or shortly after, clinical contacts. “Attachment” documents contain formal clinical correspondence, most typically between the reviewing member of staff and the referring clinician (for example, the patient’s primary care physician).

Text source	Number of times pain terms matched within the documents
Event	1,063,523
Attachments	297,538
CAMHS* Event	36,857
Discharge Notification Summary	13,175

Table 5.2. Common sources of text within whole of CRIS and the count of documents with matched pain terms within each of these sources

**Child and Adolescent Mental Health Services*

An overlap between the Attachment and Event documents that contained "pain" was examined too (Figure 5.1). In 45% of the cases, pain was mentioned in both Attachment and Events.

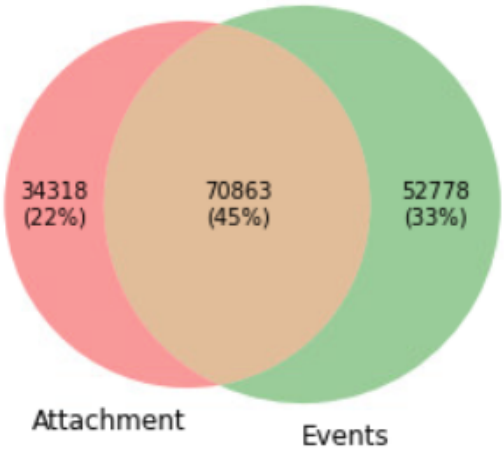


Figure 5.1. Overlap of documents mentioning "pain" between the Attachment and Event tables

To identify documents within these tables that might contain mentions of pain, a lexicon of pain terms was used. This lexicon was developed by combining multiple data sources, as described in full in (Chaturvedi et al., 2021), and Chapter 3. The lexicon consists of 382 unique pain-related terms. Since running a query on a database with such a large number of terms would be computationally expensive, the list of terms was generalised using wildcards (%), such as %pain% to capture concepts like backpain, pains, %ache for headache, and so on. After creation of wildcards, 35 unique extraction terms were used in the query. Some of these terms are shown in Table 5.3. The intention was that these limited keywords would capture all the terms within the lexicon, albeit at the risk of a lower precision than the lexicon itself. This approach led to some false positives such as paint, painting, spain for the wildcard word %pain% and attached/attaches for %ache%. Ache was modified to multiple wildcard terms such as %ache, %aches, achin%, in order to avoid picking up some of the common false positive terms. If picked up, such false positives were eliminated during the manual annotation stage. Words that could not be converted into wildcards were used in their full form, such as ‘mittelschmerz’, ‘lumbago’, ‘migraine’.

Pain word	**Word with wildcard (%)**	**Example words**
pain	%pain%	*pains, painful*
ache	%ache	*headache, backache*
ache	%aches	*Headaches, aches*
sore	sore%	*soreness, sores*
algesia	%algesi%	*analgesia, analgesic*
algia	%algia%	*proctalgia, neuralgias*

burn	%burn%	*heartburn, burns, burning*
colic	colic%	*colicky pain*
cramp	cramp%	*cramps, cramping*
dynia	%dynia%	*Allodynia, glossodynia*
hurt	hurt%	*hurts, hurting*
rheumatic	rheumati%	*rheumatic, rheumatism*
sciatic	sciati%	*sciatic, sciatica*
spasm	spasm%	*Spasms, spasmic*
tender	tender%	*tenderness*

Table 5.3. Pain words with corresponding wildcards and examples

SQL (Structured Query Language) was used to conduct the data extraction. No diagnosis or time filters were applied during this extraction. The SQL queries used in this extraction can be found on GitHub[13]. The final extraction followed the guidelines determined from the sample size calculations in Section 4.2 of Chapter 4, and went beyond the recommended minimum of 1,000 documents, achieving a total of 5,644 gold standard annotations.

5.4.4 Annotation Process

A small sample of 50 documents was extracted to examine the different contexts in which pain is mentioned. This was used to initiate the development of annotation guidelines. These

[13] https://github.com/jayachaturvedi/pain_in_mental_health

guidelines were drafted to ensure consistent annotation by multiple annotators. Upon extraction, these documents were pre-annotated with pain terms (labelled as a mention of pain, as seen italicised in Table 5.3) from the lexicon and loaded into an annotation tool, MedCAT (Kraljevic et al., 2021), for manual verification and annotation of these mentions of pain. Three medical students were employed to manually verify these annotations as well as add any associated attributes (features associated with the labelled text) based on the context around the mention of pain. These attributes are described in more detail in the upcoming paragraphs.

The first rounds of annotation, which were for the purposes of refining the annotation guidelines and training the annotators, consisted of 4 rounds. 200 documents were provided to each annotator. This number was chosen based on the time taken by the annotators. 200 documents allowed for a quick turnaround and revisions of the guidelines where required. The purpose of this was to ensure all the annotators were in agreement. At the end of each round, they provided feedback on some common false positives or ambiguous mentions, and any disagreements were discussed. Updates were made to the annotation guidelines accordingly. Inter-annotator agreements were calculated after each round of annotations. Once satisfactory inter-annotator agreement was achieved, the main annotation process commenced where each annotator was given separate sets of documents to annotate.

The annotation process (Figure 5.2) displays the steps followed by the annotators.

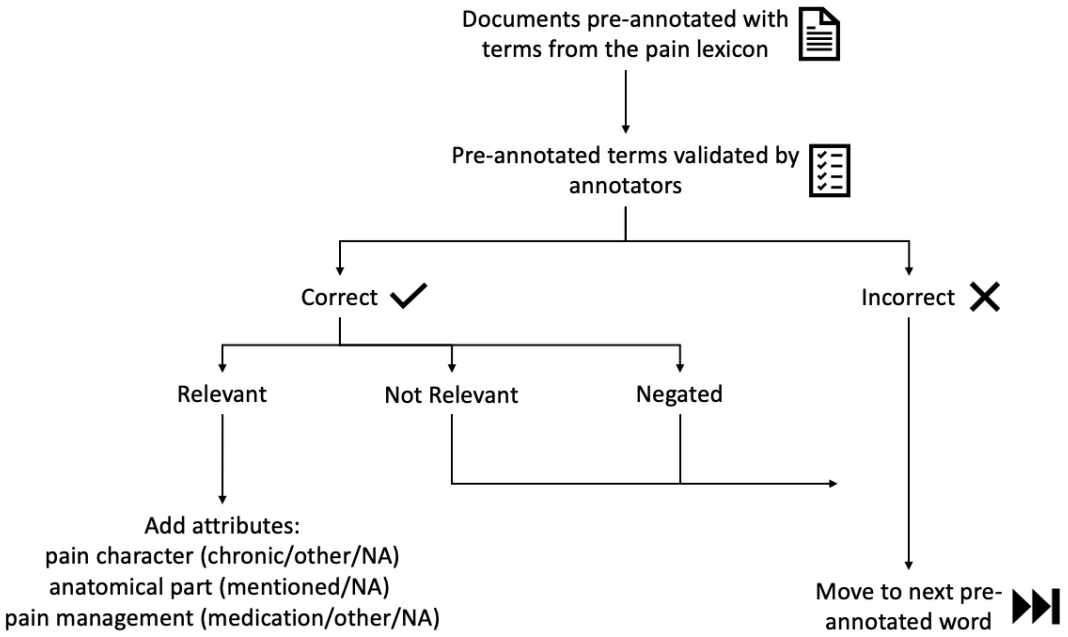


Figure 5.2. Annotation process

This flow diagram shows the annotation process followed by the medical students. They were presented with documents that were pre-annotated with terms from the pain lexicon, and they reviewed each such term, marking them as correct or incorrect. If incorrect, the annotation tool moved them to the next pre-annotated word. If correct, they further marked if the term was relevant, not relevant, or negated. If relevant, they added additional attributes. In the other two instances, they moved to the next pre-annotated word.

Annotations were marked as correct if the pre-annotated pain-related mention was in fact a mention of pain in the medical sense of the word. Mentions that were not related to human pain would be marked as incorrect, such as "...burn marks on the door" or "burning incense" for the pain-related term "burn" since the mention would be in relation to an inanimate object.

Correct mentions were labelled as "Relevant" if they were referencing pain in a medical context and it was the patient in question who was experiencing the pain. Some examples of mentions of pain that would be marked as "Not Relevant" were mentions referencing someone else's pain, such as "his mother was always in pain", or metaphorical/hypothetical mentions such as "fear of pain in the future" or the English phrase "sticking out like a sore thumb" for the pain-related term "sore". Mentions were marked as "Negated" if they

referenced absence of pain, such as “she was not in pain”, “no pain reported”, or “he does not complain of headaches”.

Relevant mentions have three further potential attributes - anatomy, pain character, and pain management.

If a mention of pain referenced a particular body part, such as “headache” indicating head, or “chronic back pain” indicating back, these were annotated as “anatomy mentioned”. If anatomy was not mentioned, the attribute defaulted to “NA”. The MedCAT annotation tool linked the annotated terms with their unique IDs from SNOMED CT which allowed for aggregation of the actual body parts at later stages of the project.

If the pain character was referenced, it was annotated as “chronic” if the character mentioned was chronic, such as “chronic back pain” or “chronic pain”, and “other” if it was any other character of pain such as “shooting pain...”, or “throbbing ache”. If pain character was not mentioned, the attribute defaulted to “NA”.

If pain management measures were indicated around the mention of pain, these were annotated as “medication” if there was reference to pain killers or other medications, or “other” for mentions like physiotherapy, pain clinic, or massage. If pain management measures were not mentioned, the attribute defaulted to “NA”.

Some examples for annotations are listed in Table 5.4.

Sentence	keyword	Correct	Relevant	Anatomy	Pain character	Pain management

He likes *burning* things..	burn	no	NA	NA	NA	NA
She *pain*ted a picture of the situation	pain	no	NA	NA	NA	NA
She is in *constant pain*	pain	yes	yes	NA	other	NA
He suffers from *severe headaches*	ache	yes	yes	mentioned	other	NA
He is not on *painkillers*	pain	yes	negated	NA	NA	medication
Afraid I will be in *pain* if surgery is unsuccessful	pain	yes	no	NA	NA	NA

Table 5.4. Examples of annotations

This table provides some example spans of sentences and how they might be annotated based on the annotation guidelines.

The annotations that were made during the training rounds went through a process of adjudication where a final annotation was chosen from the double annotated mentions based

on the latest iteration of the annotation guidelines. These adjudicated annotations, along with the main annotations, make the final corpus of annotated mentions. Upon completion of the annotation process, the prevalence of the different labels were examined. The final annotation guidelines have been made openly available for use by other researchers on similar projects[14].

5.5 Results

5.5.1 Annotation Process

Four rounds of training annotation were conducted to achieve satisfactory inter-annotator agreement. Each annotation round consisted of about 200 documents. The corresponding number of annotations and inter-annotator agreements are summarised in Table 5.5 and Figure 5.3.

Annotation round	No. of documents	No. of annotations	Cohen's Kappa	F1 score	Accuracy
Round 1	195	181	0.77	0.90	81%
Round 2	195	205	0.83	0.92	84%
Round 3	200	246	0.87	0.94	82%
Round 4	200	297	0.88	0.94	90%

Table 5.5. Summary of the overall annotation rounds.

Agreement was calculated using the Scikit-learn accuracy and Cohen's kappa function (Pedregosa et al., 2011) on the annotations. This table summarises details from the four

[14] https://github.com/jayachaturvedi/pain_in_mental_health

annotation rounds, detailing the number of documents and annotations per round, as well as metrics such as Cohen's Kappa, F1 score and Accuracy.

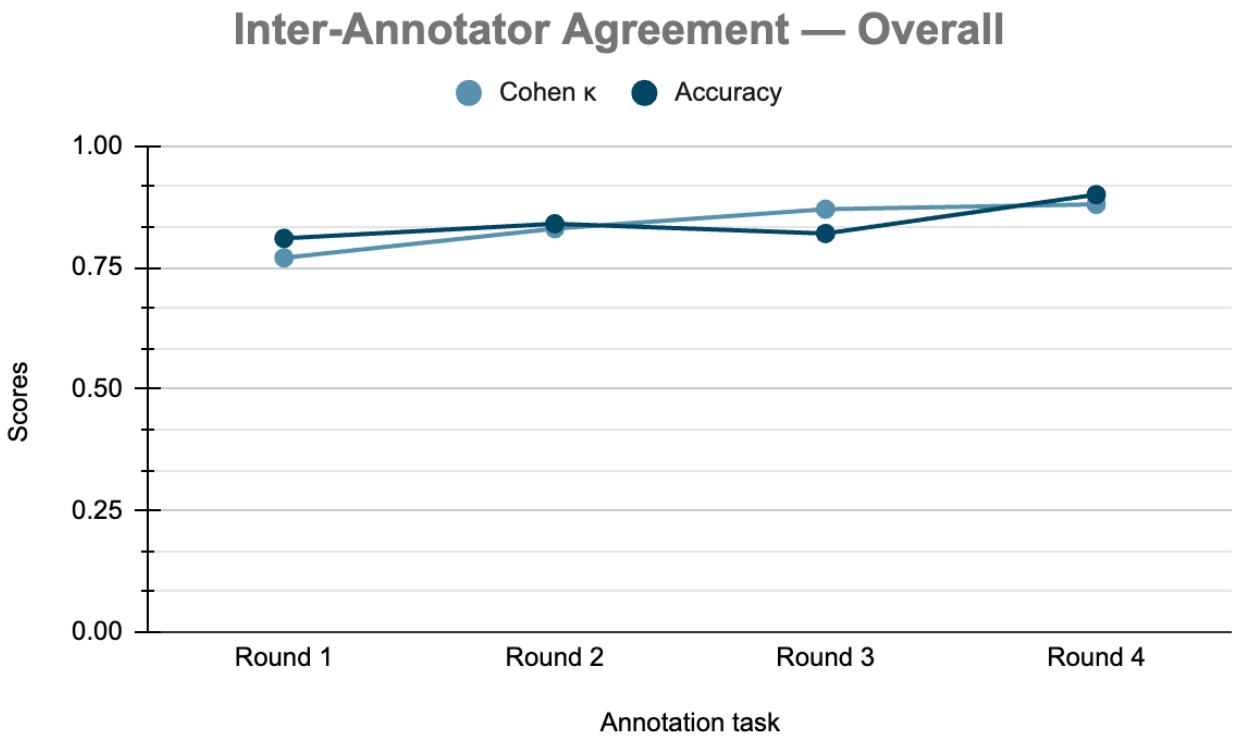


Figure 5.3. Overall Accuracy and Cohen's Kappa scores for the inter-annotator agreements across the four annotation rounds

Agreements for each attribute gradually levelled out with each round too, as displayed in Table 5.6 and Figure 5.4, with the best IAA being achieved after 3-4 rounds. This indicates that it would be unlikely for the IAA to further improve with additional rounds. These scores reflect the success of the discussions conducted with all the annotators to review any disagreements and improve the guidelines after every round of annotations.

Annotation round	**Relevant**	**Anatomy**	**Pain Character**	**Pain Management**
First	0.86	0.67	0.73	0.80

Second	0.82	0.71	0.88	0.90
Third	0.86	0.82	0.90	0.94
Fourth	0.89	0.81	0.93	0.93

Table 5.6. Summary of the inter-annotator agreement (Cohen's Kappa) on attributes for the different annotation rounds.

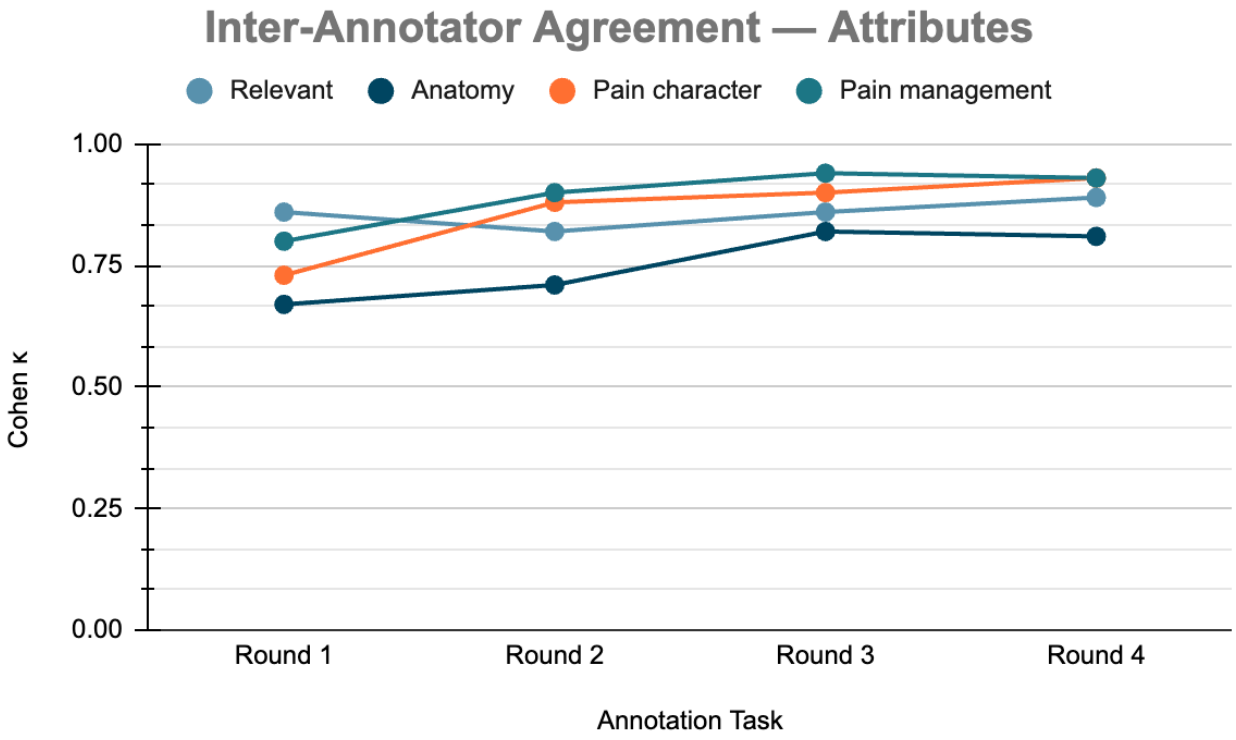


Figure 5.4. Inter-annotator agreement for the different pain attributes (Cohen's kappa) during the different annotation rounds.

A total of 5,644 annotations were collected from 1,985 documents (723 patients) (summarised in Table 5.7, Figure 5.5 and Table 5.8). This includes the adjudicated annotations from the first four training rounds where annotators double annotated the documents. The objective was to obtain a minimum of 975 annotations based on sample size calculations conducted following an approach proposed by Negida et al. (2019), but obtain more (as many as possible) if time permitted. The calculations are outlined in Appendix 1.

Total number of patients whose documents were annotated	723	
Total number of documents annotated	1,985	
Average number of words per document	1,026	
Average number of characters per document	6,474	
	Per patient	Per document
Mean/Average number of annotations	8	3
Maximum number of annotations	920	84
Minimum number of annotations	1	1

Table 5.7. Document Summary

This table summarises the total number of documents that were annotated, providing further details on the average number of words and characters per document, as well as the mean, minimum and maximum number of annotations per document and per patient.

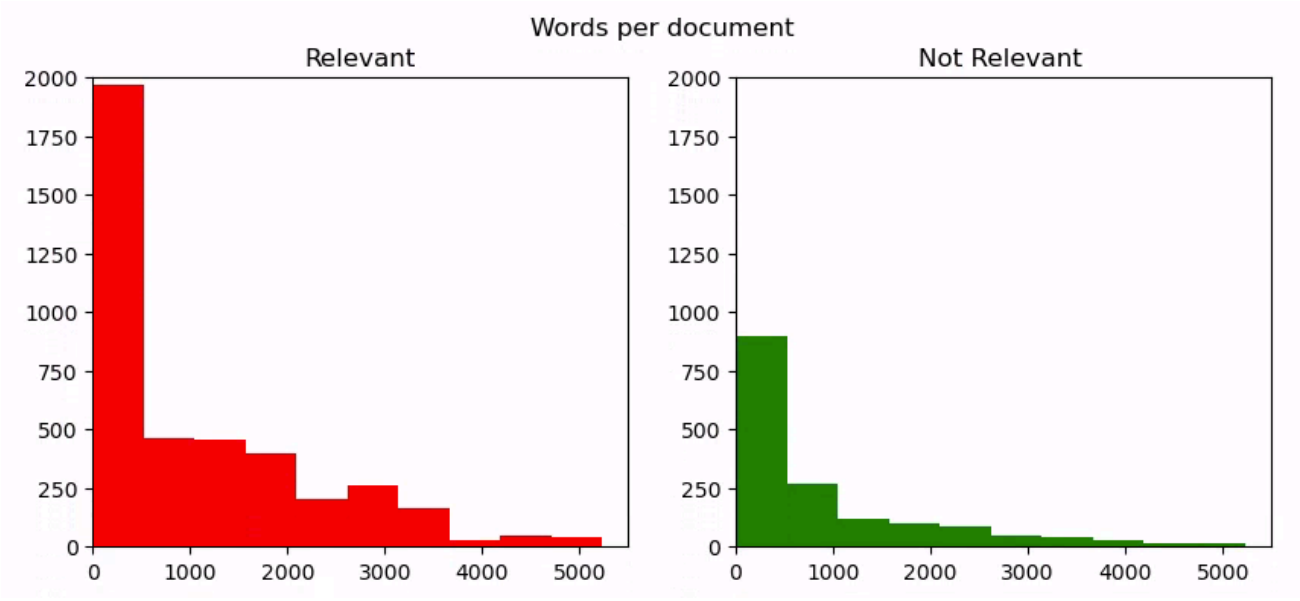


Figure 5.5. Distribution of words per document for each class

These graphs compare the distribution of words per document for both classes – relevant and not relevant.

The demographic distributions of the annotation cohort were compared to that of the CRIS population i.e., all the patients within the CRIS database, and is shown in Table 5.8.

	Annotation cohort	CRIS population
Age		
<=20	7%	17%
21 – 40	23%	36%
41 – 60	20%	30%
61-80	40%	11%
>80	10%	8%
Gender		
Male	61%	50%
Female	39%	50%
Ethnicity		
White	66%	68%
Mixed	1%	1%
Black	18%	22%
Asian	4%	4%

Other ethnic group	7%	5%

Table 5.8. Patient summary comparing the annotation cohort to the CRIS population in 2009 (Stewart et al., 2009b)

Most annotations (33%) were from patients who had a primary diagnosis of mood disorders (ICD-10-chapter F30-39) (Table 5.9).

ICD-10 chapter	**# Annotations (%)**	**CRIS Population**
Mood disorders (F30-39)	1,858 (33)	12,756 (16)
Anxiety & other non-psychotic mental disorders (F40-49)	1,122 (20)	7,105 (9)
Schizophrenia & other non-mood psychotic disorders (F20-29)	786 (14)	8,158 (10)
Mental disorder due to known physiological condition (F01-09)	460 (8)	6,414 (8)
Mental disorder, not otherwise specified	327 (6)	4,570 (6)
Mental & behavioural disorders due to substance use (F10-19)	311 (6)	7,749 (10)
Misc. (other examination or no diagnosis)	222 (4)	4,507 (6)
Person with feared complaint in whom no diagnosis is made (Z71.1)	186 (3)	-

Developmental disorders (F80-89)	104 (2)	1,541 (2)
Behavioural (F50-59)	103 (2)	1,504 (2)
Behavioural & emotional disorders, childhood onset (F90-98)	85 (2)	3,796 (5)
Personality disorder (F60-69)	59 (1)	1,291 (2)
Intellectual disabilities (F70-79)	21 (<1)	942 (1)
Total	**5,644**	**79,891**

Table 5.9. Diagnosis summary of annotations, compared to the CRIS population in 2009 (Stewart et al., 2009b)

After a few rounds of annotations of separate documents by the annotators, another inter-annotator agreement check was done to ensure there was still good agreement between the annotators. Cohen's Kappa score for the overall annotations stayed at 0.88 and accuracy went up to 92%, as compared to the scores in Table 5.5.

While the majority of the instances were straightforward to interpret as relevant and mentioning one or more of the attributes, some instances caused disagreements between the annotators, such as mentions of "period pain" and whether this should be considered relevant and a character of pain, since period pain has distinct characteristics, or whether it should be annotated as "relevant" with anatomy mentioned. It was decided that such an instance would be classified as "relevant" pain with pain character labelled as "other". Instances such as "…causing him pain" have been mentioned in situations indicating physical pain, such as "..suffering from arthritis for 20 years which is constantly causing him pain", or referencing emotional pain such as "..despite causing her a lot of pain, she returns to him". It was

important to consider the context of these mentions to decide whether they were physical or emotional mentions of pain. The issue of uncertainty also caused disagreement between annotators, where mentions of pain were followed or preceded by a question mark, such as “migraine?” or “?migraine”. Such instances were marked as “not relevant” since they were inconclusive. Ambiguous mentions such as “ongoing back pain” with no information on the time periods for the ongoing pain made it difficult to assess whether some mentions referred to chronic pain or not. Due to this, only instances which explicitly mentioned “chronic” were labelled as chronic pain.

5.5.2 Distributions of the pain attributes

Upon completion of all annotation rounds, the distributions of the various categories of annotations were summarised. The majority of the pain annotations were labelled as “Relevant” (72%) (Figure 5.6).

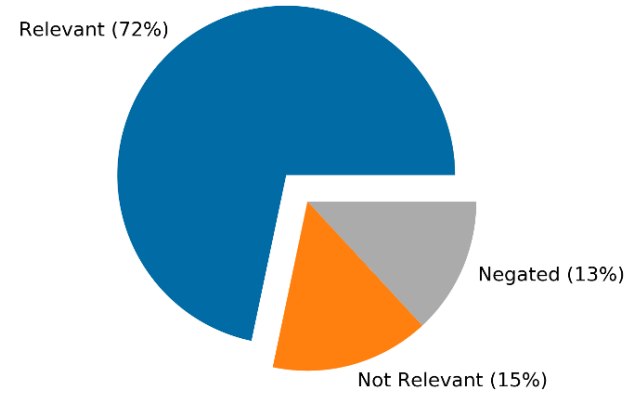


Figure 5.6. Distribution of annotations as Relevant/ Not Relevant/ Negated
This pie chart shows the distribution of the three different classes of annotated data, with the bulk belonging to the relevant class.

Amongst the relevant annotations, more than half had anatomy mentioned (Figure 5.7).

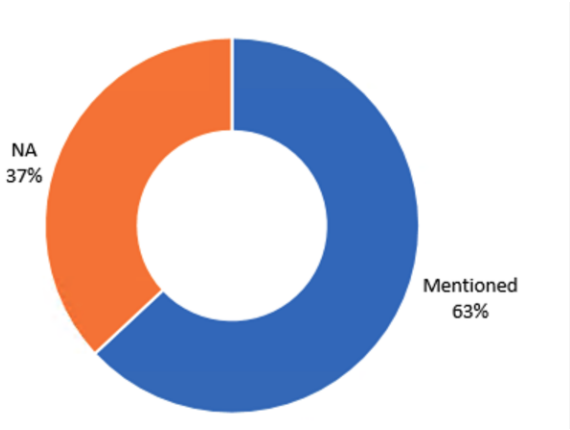


Figure 5.7. Distribution of mentions of anatomical regions within annotations
This chart shows the distribution of annotated data where the relevant mentions had anatomy mentioned, and where they did not.

Amongst the annotations with mentioned anatomical parts, the top 10 most common anatomical regions affected are shown in Figure 5.8, with chest being the most common one, followed by abdomen (including pelvis) and back. These anatomical locations were determined by linking the annotated spans to a medical terminology, SNOMED CT (Stearns et al., 2001), through the annotation tool MedCAT, which allowed for aggregation of the anatomical mentions.

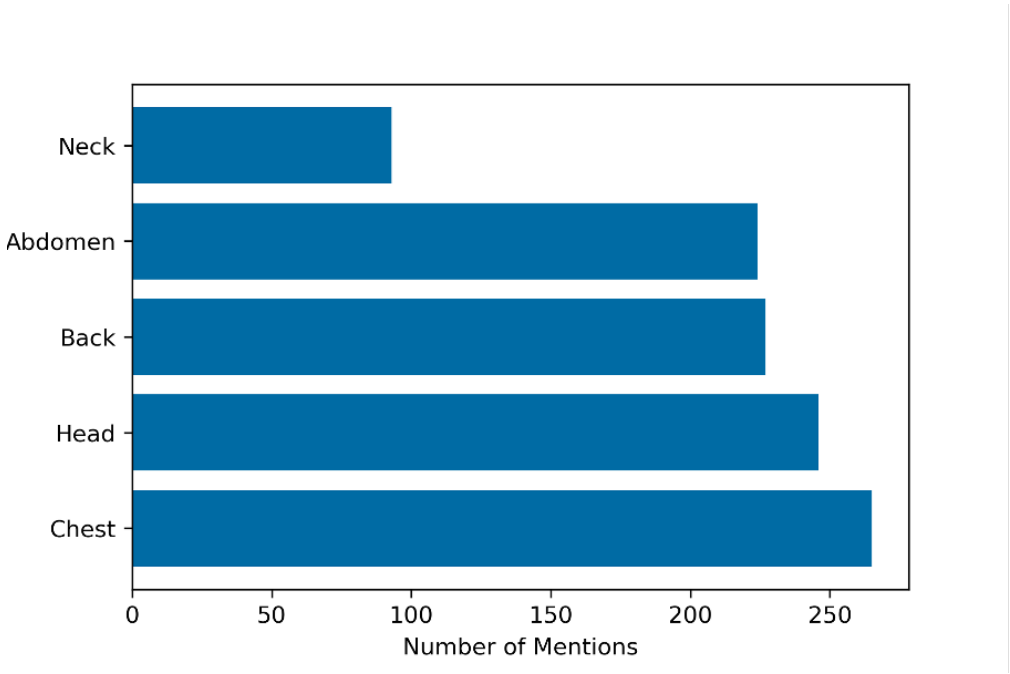


Figure 5.8. Top five most common anatomical regions within the annotations that have anatomy mentioned (n=2,540)

Similarly, amongst annotations where the pain character was mentioned as “chronic” or “other”, the majority (84%) fall under “NA” i.e., pain character was not mentioned. Apart from “NA”, “Other” was mentioned more frequently than “Chronic” (Figure 5.9).

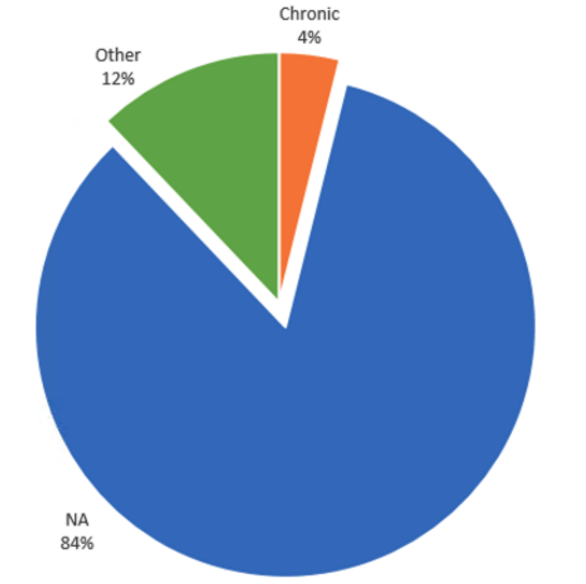


Figure 5.9. Distribution of pain character annotations

This pie chart shows the distributions of the pain character attribute for the annotations labelled as relevant. Majority did not have any pain character mentioned i.e., NA.

Amongst the annotations about pain character, chronic is most frequent, followed by burning (Figure 5.10).

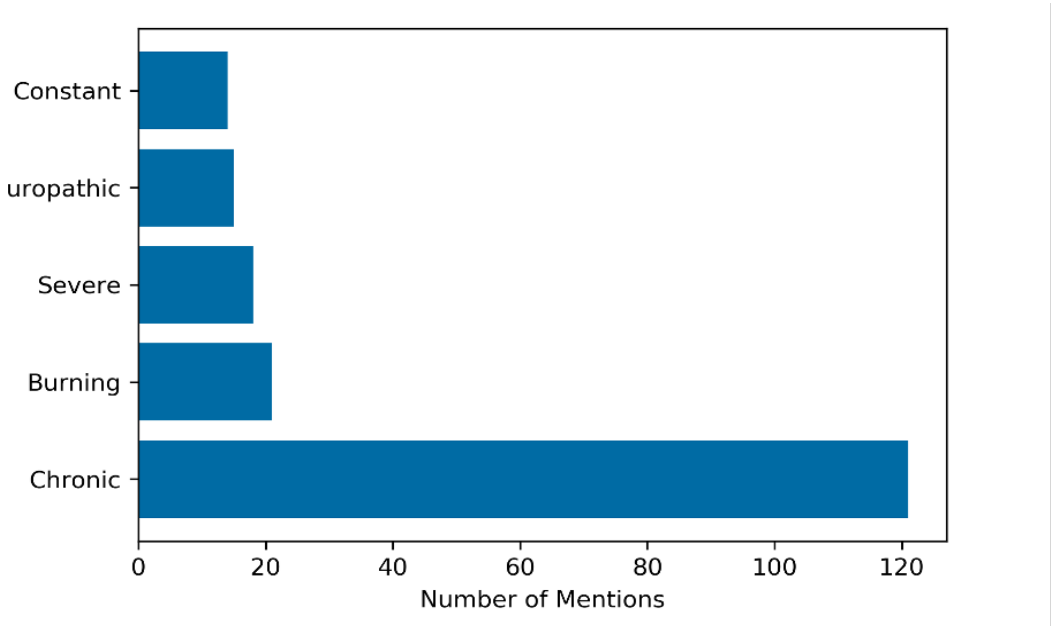


Figure 5.10. Top five pain characters mentioned within annotations marked as chronic and other (n=644)

Pain management attributes followed a similar trend where the majority were "NA" i.e., nothing about pain management was mentioned with the annotation (85%). Apart from there, Medication was mentioned more frequently than other measures (Figure 5.11).

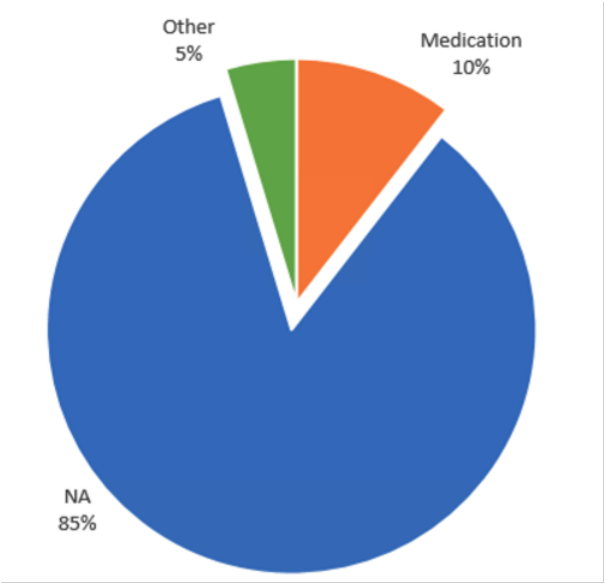


Figure 5.11. Distribution of pain management annotations
This pie chart shows the distributions of the pain management attribute for the annotations labelled as relevant. Majority did not have any pain character mentioned i.e., NA.

The “Other” pain management annotations included measures like referral to pain clinic, and physiotherapy (Figure 5.12).

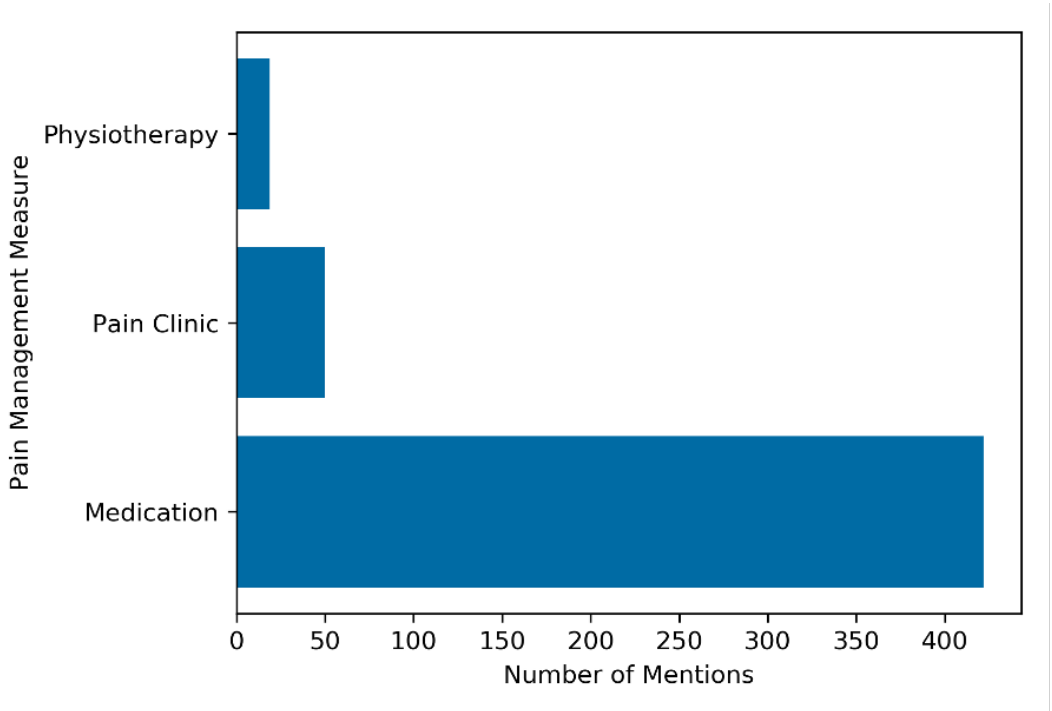


Figure 5.12. Top three pain management measures amongst the annotations marked as "Other"

The most commonly annotated concepts within the documents were "pain" (2,341 instances), "headache" (247 instances), and "painful" (206 instances).

5.6 Discussion and Conclusions

The purpose of this research was to extract mentions of pain from mental health EHRs for use in research on pain and mental health. In order to achieve this, a lexicon of pain terms was used to identify documents that contained mentions of pain and related words. These documents were then manually annotated for whether the mentions were relevant, and if so, additional attributes were labelled.

The development of this corpus has highlighted the ambiguous nature of pain, especially in mental health records, and how it could be mentioned in a variety of contextual situations. Despite this, the use of pain terms from the lexicon and achievement of good inter-annotator agreement has allowed for development of a corpus that is of good quality for use in further downstream tasks. Achievement of good inter-annotator agreement was made possible due to the methodological approach undertaken where the annotators annotated sets of 200 documents at a time and discussed any issues and disagreements before moving on to more documents. As mentioned in Section 5.5, a variety of situations caused disagreements amongst the annotators. The disagreements highlighted the importance of context around the mentions of pain. Any decisions made on such examples have been stated in the annotation guidelines that were used in the development of this corpus and are important to bear in mind when developing any machine learning algorithms. The size of this corpus and the class proportions of 72/28 (relevant/not relevant + negated) are sufficiently large for use in

development of various NLP applications (Beleites et al., 2013; Figueroa et al., 2012). Since there is some imbalance between the classes, favouring the “Relevant” class, it is important to bear this in mind to ensure that any application built using this corpus performs better than a baseline of 72% accuracy. Other means of counteracting the imbalance, such as resampling the data (oversampling the minority class or under-sampling the majority class) and SMOTE (Synthetic Minority Oversampling TEchnique) (Blagus & Lusa, 2013), can be employed as well.

A strength of this research is that it has helped to better understand how pain is mentioned within mental health EHRs, what kind of information is generally mentioned around pain in such a data source and scope the potential for the use of such information from free-text in further research. It is interesting to note that the majority of mentions of pain within these documents are relevant, and more than half of these contain information on anatomical location. Chest pain and headaches were the most frequent anatomical locations mentioned. A small portion of these relevant mentions (16%) also contained information on the pain character and any pain management measures that might have been mentioned within the sentence. Where pain management measures were mentioned, they were mostly medications such as pain killers. Mood disorders (ICD-10 chapters F30-39) were the most common primary diagnosis within the cohort (33%). This is because of the frequency of mood disorders within the CRIS database where they are the second most common primary diagnosis group (Stewart et al., 2009).

The development of this corpus is promising for future work where these annotations will be used to build an NLP application to automatically classify mentions of pain as containing anatomical location or not. This will allow for extraction of data at a larger scale with this information, so further analysis and epidemiological studies can be conducted to better

understand what body parts are commonly affected within different mental health diagnoses, and how pain experiences might differ between different diagnoses groups. There is potential to answer many more research questions around pain and mental health, and this approach will unlock the data required to do so. There are plans to link this data with primary care records (Lambeth DataNet (N H S, 2021b)), which will further improve the potential to answer critical research questions. While this corpus is not publicly available, access can be requested by contacting the CRIS administrator at cris.administrator@slam.nhs.uk

5.7 Authors' Contributions

The idea was conceived by JC, AR, RS and SV. NC, LM and VV conducted the annotations and provided insight into how pain was mentioned within the clinical notes. JC conducted the data analysis and drafted the manuscript. AR, RS and SV provided guidance in the design and interpretation of results. All authors commented on drafts of the manuscript and approved the final version.

5.8 Funding

AR was funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. RS and AR are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RS is additionally part-funded by the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust, and by the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386). JC was

supported by the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. 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 and King's College London. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.9 Declaration of Competing Interest

RS declares research support in the last 3 years received from Janssen, GSK and Takeda. The authors confirm that they have no known financial or interpersonal conflicts that would have an impact on the research presented in this study.

5.10 Appendix

Sample size calculation

We use the following method to calculate sample size (Negida et al., 2019), based on sensitivity and specificity of a similar algorithm from another study. However, we realise that this method is not ideal, and led to the simulations carried out in Chapter 4.

A study conducted by Fernandes et al. (2018) has been used in this instance as it uses the same dataset as the current project (CRIS) and applies machine learning and rules-based methods to a text classifier (Fernandes et al., 2018).

Sensitivity and Specificity from the study by Fernandes et al. (2018): “Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing”:

True Positive (TP) = 381

False Positive (FP) = 7

True Negative (TN) = 33

False Negative (FN) = 79

Sensitivity (Recall) = 98.2%

Specificity = TN/TN+FP = 82.5%

Calculation:

Z = normal distribution value = 1.96 corresponds to 95% CI

W = maximum acceptable width of the 95% confidence interval = set at 10%

Sensitivity = 98.2% from above

Specificity = 82.5% from above

$$TP + FN = Z^2 * \frac{sensitivity\,(1 - sensitivity)}{W^2}$$

$$= 1.96^2 * \frac{0.982\,(1 - 0.982)}{0.10^2} = 6.7904$$

$$TN + FP = Z^2 * \frac{specificity\,(1 - specificity)}{W^2}$$

$$= 1.96^2 * \frac{0.825\,(1 - 0.825)}{0.10^2}$$

$$= 55.4631$$

Sample size N required for sensitivity:

$$= \frac{TP + FN}{P}$$

where P is the prevalence rate and stated at 5% in the same study

$$= \frac{6.7904}{0.05}$$

$$= 135.808$$

Sample size N required for specificity:

$$= \frac{TP + FN}{1 - P}$$

$$= \frac{55.4631}{1 - 0.05}$$

$$= 58.3821$$

Therefore, total size = 136 + 59 = 195 mentions

This is for evaluation of a model already built, so can be assumed to be 20% of the dataset.

Therefore, main dataset sample = 195 x 5 = 975 mentions or 490 documents (with an average of ~3 mentions per document)

CHAPTER 6: Building a Classifier

6.1 Foreword

Following on from the creation of the gold standard annotations (Chapter 5), this chapter describes my subsequent utilisation of these annotations for building the models to classify sentences as containing relevant mentions of physical pain or not.

This work was peer-reviewed and has been accepted at the MEDINFO 2023[15] conference, presented by the Australasian Institute of Digital Health (AIDH) on behalf of the International Medical Informatics Association (IMIA). I contributed as the first author, designed the research, developed the classifier models, evaluated their performances, and drafted the manuscript. This paper was awarded the best student paper at the conference, and is available at:

Chaturvedi, J., Velupillai, S., Stewart, R., and Roberts, A. (2024). Identifying mentions of pain in mental health records text: A natural language processing approach. Studies in Health Technology and Informatics, 310, 695–699. https://doi.org/10.3233/

Furthermore, I had the opportunity to compose a blog post about this paper for the National Institute of Health Research Biomedical Research Centre (NIHR BRC). This blog post received extensive promotion across various social media platforms and is accessible at the following link: https://www.maudsleybrc.nihr.ac.uk/posts/2023/august/identifying-mentions-of-pain-in-mental-health-records-text-a-natural-language-processing-approach/

[15] https://medinfo2023.org/

The following sub-sections of this chapter reproduce the preprint of the paper, with some minor formatting adjustments to keep it in line with the format of the thesis. The content itself has not been altered.

However, the paper does not include the training parameters and performance metrics for the Random Forest classifier or another classifier built to identify whether anatomy was mentioned amongst the relevant pain sentences. Given the recent popularity of GPT models, a GPT-2 model was also trained. In addition to information about these classifiers, the paper does not include much detail about how the class imbalance in the training dataset was handled. Details about these have, therefore, been added separately (Section 6.9) at the end of this chapter.

Identifying Mentions of Pain in Mental Health Records Text: A Natural Language Processing Approach

Jaya CHATURVEDI[a], Sumithra VELUPILLAI[a], Robert STEWART[a,b,c] and Angus ROBERTS[a, b]

[a] *Institute of Psychiatry, Psychology and Neurosciences, King's College London*

[b] *Health Data Research UK*

[c] *South London and Maudsley Biomedical Research Centre, London, United Kingdom*

ORCiD ID: Jaya Chaturvedi https://orcid.org/0000-0002-6359-9853

Abstract. Pain is a common reason for accessing healthcare resources and is a growing area of research, especially in its overlap with mental health. Mental health electronic health records are a good data source to study this overlap. However, much information on pain is held in the free text of these records, where mentions of pain present a unique natural language processing problem due to its ambiguous nature. This project uses data from an anonymised mental health electronic health records database. The data are used to train a machine learning based classification algorithm to classify sentences as discussing patient pain or not. This will facilitate the extraction of relevant pain information from large databases, and the use of such outputs for further studies on pain and mental health. 1,985 documents were manually annotated for creation of gold standard training data (800 of these were triple annotated), which was used to train three commonly used classification algorithms. The best performing model achieved an F1-score of 0.98 (95% CI 0.98-0.99).

Keywords. Natural Language Processing, Electronic Health Records, Pain, Mental Health, Transformers.

6.2 Introduction

Pain is defined as an unpleasant sensory and emotional experience, and is influenced by a variety of biological, psychological, and social factors (International Association for the Study of Pain, 2020). Pain is very subjective in nature and best described verbally by the person experiencing it (International Association for the Study of Pain, 2020). Pain is a global healthcare problem, and consequently a growing area of research. A high co-occurrence of pain and mental health disorders has been established and known to be linked to increased disability and impairment (Vinall et al., 2016). Pain is a common reason for people to access healthcare facilities, thereby making electronic health records (EHR) a potential source for information on pain (Motov & Khan, 2008).

EHRs are longitudinal compilations of electronic data pertaining to a person's medical history or healthcare (Institute of Medicine (US) Committee on Data Standards for Patient Safety, 2003). They have been increasingly used in research as they provide the opportunity to explore patient symptoms and findings from both structured fields (such as demographics, diagnosis, lab test results etc.) and unstructured fields (such as clinical notes, discharge summaries, referral letters etc). Previous research has attempted the identification of patient pain experiences by combining individual elements such as pain scores, prescription medications, and billing codes (Tian et al., 2013). Since pain is not well recorded in these structured fields, it may help to supplement this information from the structured fields with data from unstructured clinical text. These unstructured narratives within clinical notes help provide context to the patient’s pain since their individual and self-reported experiences are generally documented in free text (Carlson & Hooten, 2020).

To extract information from textual sources we can utilise natural language processing (NLP). NLP is a subfield of artificial intelligence used to leverage rich textual sources for information extraction and retrieval (Liu et al., 2011) by transforming natural language into computational representations for analysis (Yim et al., 2016). Due to the ambiguous and sometimes metaphorical nature of how pain is used in communication, recent advances in NLP which adopt contextual and metaphorically informed methods could contribute to a deeper comprehension of how pain affects health and the utilization of healthcare resources in the treatment of pain (Carlson & Hooten, 2020).

Due to the availability of large amounts of data from EHRs, there have been increased opportunities to use machine learning approaches on such "big data" more effectively (Koopman et al., 2015; McCowan et al., 2007; Rajkomar et al., 2018). A commonly used machine learning based NLP approach is text classification, in which labels are assigned to units of text (sentences/paragraphs/documents). This is frequently seen in applications such as sentiment analysis and spam detection (Minaee et al., 2021). Within the healthcare domain, this can be used to classify presence or absence of features such as symptoms/diagnosis/smoking status and so on. Commonly used classification algorithms include Support Vector Machines (Garla et al., 2013; Wright et al., 2013; Yao et al., 2019) and K-Nearest Neighbours (Jindal & Taneja, 2015; Trstenjak et al., 2014; Xing & Bei, 2020). Recent state of the art approaches use embedding models and transformer-based neural network architectures (Vaswani et al., 2017), such as the bi-directional encoder representations of BERT (Devlin et al., 2018). These approaches build unsupervised language models from large general corpora, which can be transferred to other learning problems by fine tuning on smaller datasets. The BERT base model is trained on 3.3 billion words from the general domain (Wikipedia and BookCorpus) (Devlin et al., 2018), with many healthcare domain related models emerging such as PubMedBERT (Gu et al., 2022),

BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), UmlsBERT (Michalopoulos et al., 2021) and SapBERT (Liu et al., 2021) which were developed after recognition of the need for specialised models due to linguistic differences between general and biomedical text (Alsentzer et al., 2019).

This paper describes the methods undertaken to develop an NLP application for a sentence-level classification of mentions of pain within clinical text, as mentioned in Chapter 2. Due to the messy nature of the textual data within EHRs, it is challenging to split the text into sentences (lack or inconsistent use of punctuations). For this reason, sentences here are determined as 200 characters before/after a pain term. Two BERT models were trained - BERT_base and SapBERT - and compared to two conventional models - support vector machines (SVM) and K-Nearest Neighbours (KNN). The outputs from this application can be used to extract relevant pain information from large EHR databases for use in further epidemiological studies and pain related research. To the best of our knowledge, such extraction of information about pain from mental health clinical text using NLP has not been done, while similar work has recently been undertaken in a non-mental health setting (Dave et al., 2022; Naseri et al., 2023).

6.3 Methods

6.3.1 Data Source

An anonymised version of EHR data from The South London and Maudsley NHS Foundation Trust (SLaM), one of the largest mental healthcare organizations in Europe, is stored in the Clinical Record Interactive Search (CRIS) database (Stewart et al., 2009b). The infrastructure of CRIS has been described in detail with an overview of the cohort profile (Perera et al.,

2016). CRIS contains over 30 million documents, averaging 90 documents per patient (Velupillai et al., 2018). There are 23 different text sources (such as attachments, event notes, nurse assessment letters, etc.). Most of the text is contained within attachments and event notes, and so these were used as the data sources in this project.

6.3.2 Ethics and Data Access

Ethics approval for CRIS has been granted by (Oxford C Research Ethics Committee, reference 18/SC/0372). Research projects that use the CRIS database are reviewed and approved by a patient-led oversight committee (described in (*CRIS - South London and Maudsley NHS Foundation Trust.*, 2022)). An opt-out model is in place for service users and is advertised in all publicity material and initiatives. Data are owned by a third party, Maudsley Biomedical Research Centre (BRC), who run the CRIS tool, providing access to anonymised data. These data can only be accessed by permitted individuals from within a secure SLaM firewall. All access is password protected using authenticated SLAM network usernames and passwords. Data exported from CRIS cannot be saved directly outside the firewall. Further technical details of the security model can be accessed on their website[16].

6.3.3 Data Extraction

Pain can be described in numerous ways, using a variety of terms. To help identify which documents in CRIS might be discussing pain, a lexicon of such pain terms was developed from a combination of pain-related terms extracted from the literature and biomedical ontologies, supplemented with additional similar terms from word embedding models. This lexicon and its development has been described in more detail in (Chaturvedi et al., 2021) and Chapter 3. Terms from this pain lexicon were used to identify documents within CRIS

[16] https://www.maudsleybrc.nihr.ac.uk/media/112184/cris-security-model.pdf

that might be discussing pain. Documents containing pain terms were extracted for further processing using SQL. No time or diagnosis filter was applied to the extraction.

6.3.4 Annotation Task

Extracted documents were used to create a corpus of text discussing patient pain by labelling, i.e., annotating spans of text as being about pain or not. Each span consisted of 200 characters before and after a pain-related term. As previously described in Chapter 5, first, a set of annotation guidelines were developed to provide rules defining when a sentence should be considered as discussing pain. These guidelines were illustrated with examples. Next, terms from the pain lexicon were highlighted in the extracted documents. This was done for easy identification of pain-related sentences within the long documents. Three medical student annotators (the same annotators from Chapter 5) read through the extracted documents considering these spans of text containing the previously highlighted pain terms. Annotators labelled each span with one of three labels: *relevant*, i.e., referring to physical pain experienced by the patient; *not relevant* i.e., mentions not related to pain, not related to the patient or hypothetical and metaphorical mentions; and *negated* i.e., absence of pain. Instances that were marked as relevant pain mentions were further annotated with attributes such as pain character (chronic/other/NA), anatomy (mentioned/NA), and pain management (medication/other/NA). Annotators also marked any other spans that contained pain-related words but were not pre-annotated. Inter-annotator agreements were calculated after multiple rounds where 200 documents were annotated by all three annotators in each round. The annotation tool used for this was MedCAT (Kraljevic et al., 2021). Any disagreements were discussed, and the annotation guidelines updated. This iterative process was carried out until an inter-annotator agreement of over 0.80 (accuracy and Cohen's kappa) was achieved, as described in Table 5.4 of Chapter 5. Upon receiving satisfactory inter-annotator agreement,

each annotator was then given a separate set of documents. The documents that were annotated by all three annotators during the iteration process were adjudicated following a set of adjudication guidelines in line with the most recent version of annotation guidelines. The purpose of adjudication was so that all the annotations could be used as gold standard. The annotation and adjudication guidelines can be accessed online[17].

6.3.5 NLP application

Spans labelled by the annotators were used as gold standard training data for development of the NLP application. The annotations were split into train/test/validation sets at a proportion of 80/10/10 respectively. Pre-processing including lowercasing, removal of stopwords, and tokenisation was carried out. Four different models were trained, as detailed in Table 6.1. Two versions of BERT models were used, BERT_base and SapBERT. The parameters were chosen based on the recommendations made by Devlin et al. (2019) and models were checked for overfitting.

Model	Tokeniser	Pre-processing	Other Parameters
1. Support Vector Machine 2. K-Nearest Neighbour	NLTK	Lowercase, stopword, white space and punctuation removal, lemmatise and tokenise	Tf-Idf vectorizer Default parameters from sklearn
3. BERT_base	bert_base_uncased	Tokenise Prepend sentence with special token [CLS] and append with special token [SEP]	Epochs: 3 Batch size: 16 Optimizer: AdamW, learning rate 3e-5

[17] https://github.com/jayachaturvedi/pain_in_mental_health/blob/main/Annotation%20Guidelines%20-%20Pain%20-%20for%20github.pdf

4. SapBERT	cambridgeltl/SapBERT-from-PubMedBERT-fulltext	Pad and truncate sentence to max length 105 (default is 511)	Epochs: 4 Batch size: 16 Optimizer: AdamW, learning rate 2e-5

Table 6.1. Model specifications

This table details the parameters for the training of each model, including any pre-processing measures taken, tokenisers used, and others such as epochs, optimizers and learning rates, where applicable.

6.4 Results

6.4.1 Data Extraction

A total of 1,985 randomly selected documents from 723 patients were extracted that contained pain related keywords from the lexicon. The most common diagnosis codes for these extracted patients were Mood disorders (ICD10 chapters F30-39) (33% of patients). There was an average of 8 annotations per patient.

6.4.2 Annotations

An inter-annotator agreement of 90% (Cohen's kappa 0.88) was achieved after four rounds (each round containing 200 documents) of triple annotations. Upon completion of adjudication on the documents that had disagreements, along with annotations conducted on individual documents by the three annotators, a total of 5,644 annotations were obtained. 72% of these were marked as relevant, 15% as not-relevant, and 13% as negated. Amongst the relevant annotations, 45% had anatomy mentioned, 11% had pain character mentioned, and 10% had pain management measures mentioned. The relevant annotations were labelled as 1. The not-relevant and negated annotations were combined and labelled as 0.71% annotations were labelled as class 1 (relevant) for both training and testing data.

6.4.3 Evaluation of NLP application

A single GPU (Tesla T4) was used for training the models. All sentences labelled as 0 and 1 were used as the training data. Sentences that were not annotated were not included in the training data. K-fold validation (K=10) was carried out on the training data (80% of the data) for evaluation of the models, and 95% confidence intervals were calculated using a bootstrapping method (bootstrap sample datasets = 500). The training data (The results for each algorithm are outlined in Table 6.2. The BERT models performed better than Support Vector machine and K-Nearest Neighbour models.

Model	Precision	Recall	F1-score (average from 10-fold cross validation)
Support Vector Machine	0.86 (0.83-0.88)	0.98 (0.97-0.99)	0.91 (0.90-0.93)
K-Nearest Neighbour	0.84 (0.81-0.87)	0.91 (0.89-0.93)	0.87 (0.85-0.89)
BERT_base	0.96 (0.94-0.97)	0.98 (0.97-0.99)	0.97 (0.96-0.98)
SapBERT	0.98 (0.97-0.99)	0.99 (0.98-0.99)	0.98 (0.98-0.99)

Table 6.2. Evaluation Metrics for the 4 models, including 95% confidence intervals

Performance on the validation dataset (10% of the data) is detailed in Table 6.3

Model	Precision	Recall	F1-score
Support Vector Machine	0.89 (0.87-0.91)	0.89 (0.86-0.91)	0.88 (0.85-0.90)
K-Nearest Neighbour	0.81 (0.78-0.84)	0.82 (0.79-0.84)	0.81 (0.79-0.84)
BERT_base	0.94 (0.93-0.96)	0.94 (0.93-0.95)	0.94 (0.92-0.95)

SapBERT	0.95 (0.93-0.97)	0.94 (0.92-0.96)	0.94 (0.92-0.96)

Table 6.3. Evaluation Metrics for the 4 models on validation dataset, including 95% confidence intervals

The models were checked for overfitting. Figure 6.1 shows the training and validation loss for the SapBERT model steadily decreasing with each epoch, apart from a slight increase in the validation loss from epoch 3 to 4.

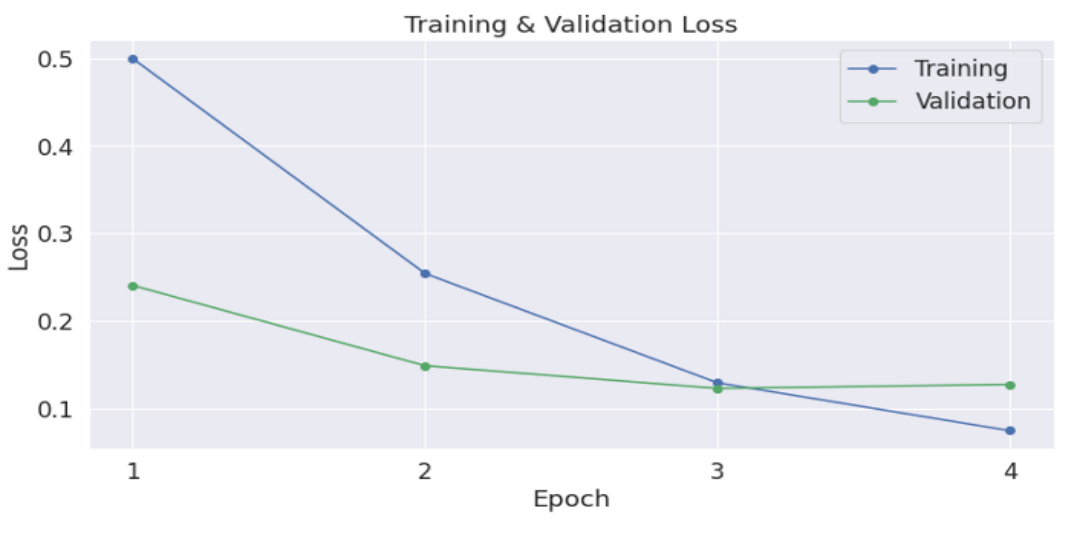


Figure 6.1. Training and validation loss – SapBERT
This graph shows the loss values per epoch, through the training and validation process for the SapBERT model.

6.4.4 Error Analysis

During the annotation process, common disagreements included when an instance could be interpreted as physical or metaphorical, such as “...causing him pain”, and hypothetical mentions such as “...she feared the pain” and “?migraine”.

After training the models, some false positives spotted during error analysis on the test data for the BERT model were instances such as “...wishing to project his pain on others”, “father’s hip pain”, “...headaches were a common adverse effect reported by the trial”. Some false negatives such as “denying symptoms other than stomach ache”, “...if pain increases”, “bruised arm is painful, no other worrying findings” were also noted.

The SapBERT model showed false negatives when there were undecipherable symbols incorporated in the text, which might have occurred during the anonymisation process of the text, as well as misspellings or conjoined words such as “dabdominal pain” and “achespainodd sensations”. False positives were instances such as “risk of potential pressure sores”.

The Support Vector Machine model showed false negatives such as “agitated in the context of pain” and shortened words such as “abdo pain”.

6.5 Discussion

Pain is very subjective and ambiguous in its description, making it hard for clinicians to code pain within structured fields of EHRs. The free-text fields within the EHR provides clinicians with the flexibility to describe the pain in the patient’s own words or based on their interpretations. The ambiguous nature of pain was highlighted during this project, especially during the annotation process where it took multiple rounds for three clinically trained annotators to agree on the meanings and interpretations of the pain mentions. Bearing this in mind, it is understandable that the classification models struggled with hypothetical and metaphorical instances. It is also important to note that this project is focused on identifying instances of physical pain, so any references to mental pain and anguish would have to be distinguished from physical pain too. This highlights the importance of context and the

necessity for the NLP models to incorporate and consider context during the classification task. This is a strength of transformer-based models such as BERT, in addition to BERT models being pretrained on large corpora prior to any fine-tuning, which could be why they performed better than SVM/KNN.

Amongst the two BERT models that were trained, SapBERT, which was pre-trained using a biomedical ontology, UMLS, performed better than BERT_base, with a slight increase in performance scores compared to BERT_base. BERT models utilise a WordPiece tokenisation algorithm where if a word does not exist within the BERT vocabulary, it divides the words into subword units by adding a prefix (such as ##) (*WordPiece Tokenization - Hugging Face Course*, 2022). There were differences in how each of the BERT models used in this project tokenised words, where SapBERT was able to tokenise clinical concepts more accurately. For example, a word like “sciatica” is split into “sci”, “##atic”, “##a” when using BERT tokenisers, where it uses a WordPiece tokenisation algorithm and adds ## when a word is not in its vocabulary (*WordPiece Tokenization - Hugging Face Course*, 2022). SapBERT was able to tokenise this as a complete word “sciatica”. This improvement in tokenisation might have impacted and improved the overall performance of the model.

As mentioned in Section 6.4.2, the gold standard annotations consisted of three classes – relevant, negated and not relevant, and the negated and not relevant classes were combined for the purposes of the classification task. However, the negated class can be used as a distinct class in future work, if required, for work that might focus on distinguishing negations with respect to pain mentions.

6.6 Conclusions

The objective of this project was to develop a machine learning based NLP application that can classify mentions of pain within clinical text as relevant or not. BERT models, which use a transformer-based machine learning technique and contextual embeddings, outperformed the other algorithms. This is a novel approach towards extracting information about pain from mental health records, leveraging the unstructured clinical notes to identify patients with relevant mentions of pain, and such cohorts of patients can then further be used in epidemiological and other pain related research with more confidence in the actual occurrence of pain when mentioned in the text.

6.7 Acknowledgements

AR is funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. RS and AR are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RS is also part-funded by: i) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King’s College Hospital NHS Foundation Trust; ii) the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386). JC is supported by the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. RS declared research support

received in the last 36 months from Janssen, GSK and Takeda. All other authors declare no other competing interests.

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 and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the 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.

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities.

This work uses data provided by patients and collected by the NHS as part of their care and support. An application for access to the Clinical Record Interactive Search (CRIS) database for this project was submitted and approved by the CRIS Oversight Committee. The authors are also grateful to Dr Aurelie Mascio for providing access to some of her Python scripts.

6.8 Class Imbalance

All the classifiers built in this project utilise the same set of gold standard annotations, as detailed in Section 5.5.2 of Chapter 6. 72% of the annotations belonged to the “relevant” class, leading to a class imbalance and a potential bias towards the majority class and low accuracy when predicting the minority class. Methods to handle class imbalance include resampling the data (oversampling the minority class or under-sampling the majority class), the SMOTE (Synthetic Minority Oversampling TEchnique) algorithm (Blagus & Lusa, 2013), and incorporation of cross-entropy loss within the training of the model. The latter has been used in this thesis. A loss function quantifies how well a model's predictions match the target values. Knowing this helps understand how well the model is doing and whether it requires further parameter adjustments to aid in better performance. Cross-entropy loss is a type of loss function widely used in machine learning which measures the difference between the predicted probabilities of the class labels and the true class labels. By incorporating weighted cross-entropy loss in the training of a model, the imbalance in class distribution is addressed by encouraging the model to pay more attention to the minority class. When a larger weight is assigned to the minority class, it helps counteract the class imbalance, leading the model to make more informed decisions for both classes (Rezaei-Dastjerdehei et al., 2020). While the SapBERT model inherently applies a multi-similarity loss to deal with class imbalance, which leverages the similarities between the positive and negative pairs and re-weights the importance based on this, such that more informative pairs receive more gradient signals during training and therefore can better use the information stored in data, the weighted cross-entropy was used in addition to this to ensure the imbalance was addressed.

Cross-entropy loss was applied to all classifiers described in the previous section. Since SapBERT performed best, a brief comparison was run between the model with and without this loss (Table 6.4) to understand the implication of incorporating it into the model.

Model	Label	Precision	Recall	F1-score
SapBERT without weighted cross-entropy loss	0	0.98 (0.96-0.99)	0.95 (0.93-0.97)	0.97 (0.95-0.99)
	1	0.98 (0.96-0.99)	0.99 (0.97-0.99)	0.99 (0.98-0.99)
	Weighted Average	0.98 (0.97-0.99)	0.99 (0.98-0.99)	0.98 (0.97-0.99)
SapBERT with weighted cross-entropy loss	0	0.92 (0.89-0.94)	0.86 (0.83-0.88)	0.89 (0.86-0.91)
	1	0.95 (0.93-0.96)	0.97 (0.95-0.98)	0.96 (0.94-0.97)
	Weighted Average	0.94 (0.92-0.95)	0.94 (0.92-0.95)	0.94 (0.92-0.95)

Table 6.4. Comparison of SapBERT model with and without cross-entropy loss, with 95% confidence intervals

While the performance metrics dropped a small amount after incorporation of cross-entropy loss, especially with class 0, there was 86% similarity between the two, i.e., both models generated the same predictions for 86% of the sentences. A sample of 2000 sentences was manually reviewed to analyse differences in predictions from the two models. Amongst these sentences, the model incorporating cross-entropy loss made the correct predictions more frequently when faced with instances that commonly caused misclassifications, such as when a person's name was mentioned but picked up as a pain term (such as Dr Sore, Mr Burn). The model with cross-entropy loss classified these instances correctly 93% of the time. Similarly, false positives such as spain and paint were correctly classified by this model 81% of the time.

6.9 Additional Classifiers

6.9.1 Random Forest Classifier

In addition to the classifiers mentioned in Section 6.3.5 of this chapter, a random forest classifier was developed. This allowed for a comparison with the random forest classifier incorporating external knowledge to be discussed in Section 7.9 of Chapter 7. The same gold standard annotations previously mentioned (in Section 5.5.1 of Chapter 5) were used for training and testing this classifier. The parameters for training the model are specified in Table 6.5.

Model	Tokeniser	Pre-processing	Other Parameters
Random Forest	NLTK	Lowercase, stopword, white space and punctuation removal, lemmatise and tokenise	Tf-Idf vectorizer Default parameters from sklearn

Table 6.5. Model specifications

This table details the parameters for the training of each model, including any pre-processing measures taken, and tokenisers used.

Similar to the other classifier models, K-fold cross-validation was carried out for evaluation of the model, and 95% confidence intervals were calculated. The results are outlined in Table 6.6.

Model	Precision	Recall	F1-score (average from 10-fold cross-validation)

Random Forest	0.86 (0.84-0.88)	0.86 (0.83-0.88)	0.85 (0.83-0.87)

Table 6.6. Evaluation Metrics (weighted average) for the Random Forest classifier model, including 95% confidence intervals

Error analysis of the test set predictions found false negatives such as negation followed by a positive instance, for example “initially denied pain but then mentioned headache..”, and misclassified instances where symbols preceded the mention of pain. These symbols were potentially introduced into the text during the anonymisation process, and not picked up by the punctuation removal in pre-processing. False positives included negated instances such as “did not appear to be in pain” or “pain Nil”, and metaphorical mentions like “painful to think about” or “pretended to have stomach ache”.

6.9.2 GPT-2 Classifier

Due to the rise in popularity of the GPT models, a classifier was also trained with the latest freely available GPT model, GPT-2 (Radford et al., 2019), to evaluate if it performs better than the other models trained thus far. The data was split into 80/20 proportions for train/test, and the parameters mentioned in Table 6.7 were used. The data split for this model is different from the ones previously described because this model was trained at a different time point, once the GPT-2 model was made publicly available, which led to minor inconsistencies in parameters, such as the data split, when compared to the other models. Due to this, the results from the GPT-2 model are not entirely comparable with the previously described model.

Model	Tokeniser	Pre-processing	Other Parameters
GPT-2	gpt2	Tokenise	Epochs: 4

	Pad and truncate sentence to max length 60 (default is 511)	Batch size: 32 Optimizer: AdamW, learning rate 2e-5

Table 6.7. Model Parameters

This table details the parameters for the training of each model, including any pre-processing measures taken, tokenisers used, and other training parameters such as epochs, optimizers and learning rates.

To ensure there was no overfitting, the training and validation losses were plotted, and a steady decline can be seen in Figure 6.2.

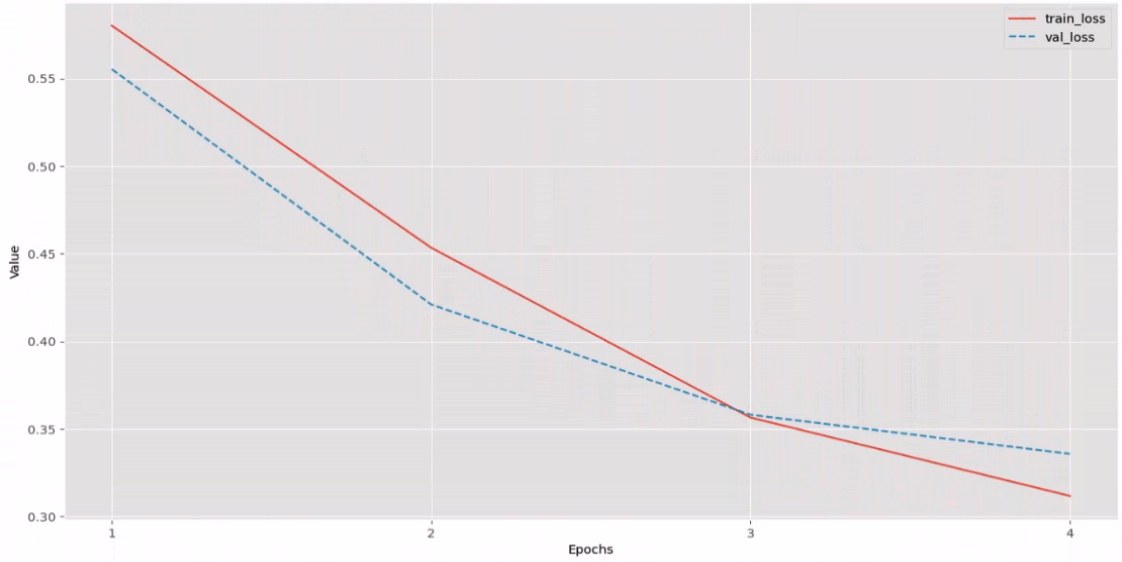


Figure 6.2. Training and validation loss

This graph shows the loss values during training and validation, per epoch, for the GPT-2 model.

Similarly, a steady increase is seen in the training and validation accuracy, which starts to plateau by epoch 4 (Figure 6.3)

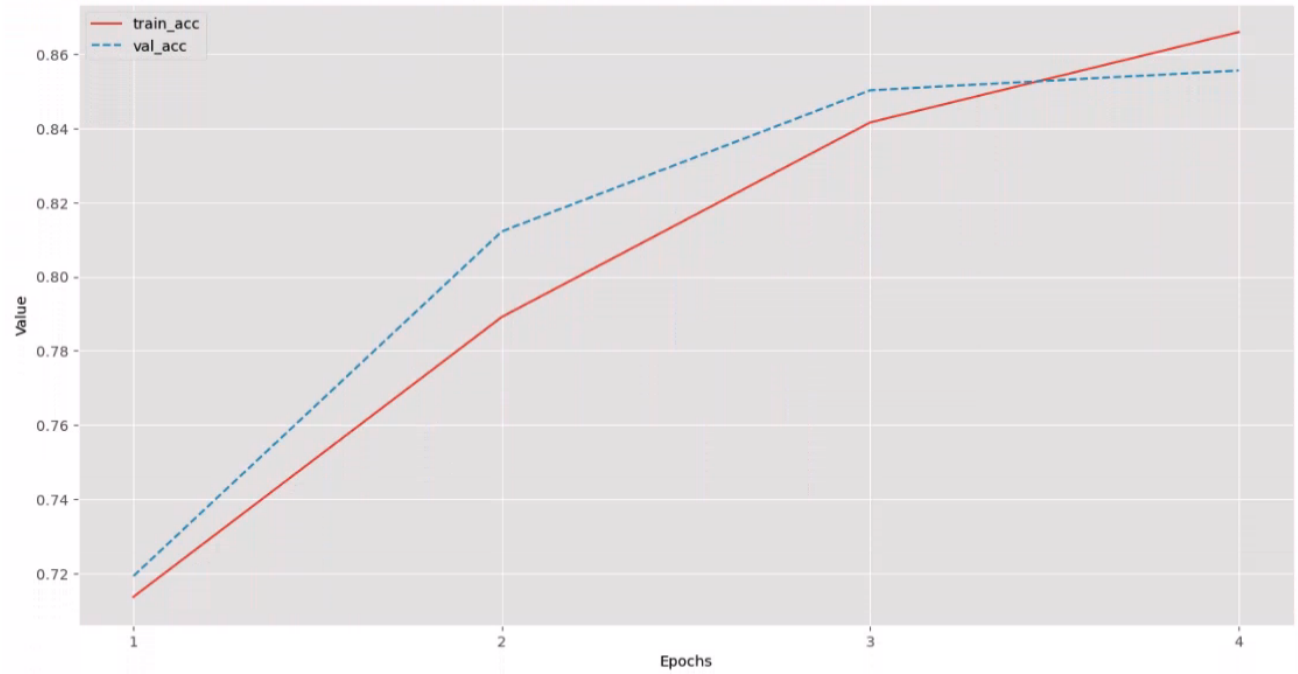


Figure 6.3. Training and validation accuracy per epoch for the GPT-2 model
This graph shows the accuracy values during training and validation, per epoch, for the GPT-2 model.

The final performance metrics are detailed in Table 6.8, as well as a normalised confusion matrix in Figure 6.4.

Model	Precision	Recall	F1-score (average from 10-fold cross-validation)
GPT-2	0.85 (0.83-0.87)	0.85 (0.83-0.87)	0.85 (0.83-0.87)

Table 6.8. Performance metrics for the GPT-2 model, including 95% confidence intervals

The normalised confusion matrix shows a majority of the classification instances as being true positive (0.93) and true negative (0.68). However, there is a higher proportion of false positives (0.32) compared to false negatives (0.07). This imbalance may stem from the nature

of GPT-2's training data, which primarily consists of generic English text including medical literature, which is significantly different from the language used in clinical settings. Furthermore, models like GPT-2 are primarily designed for tasks such as natural language understanding and text generation, rather than classification problems. As a result, they may struggle with classification tasks, specifically in the medical domain. This discrepancy highlights the potential limitations of directly applying general purpose language models to domain specific tasks. Default parameters were used for the fine-tuning of this model, and while GPT-2 demonstrates promising capabilities in capturing broad language patterns, its performance on this clinical classification task highlights the need for further refinement.

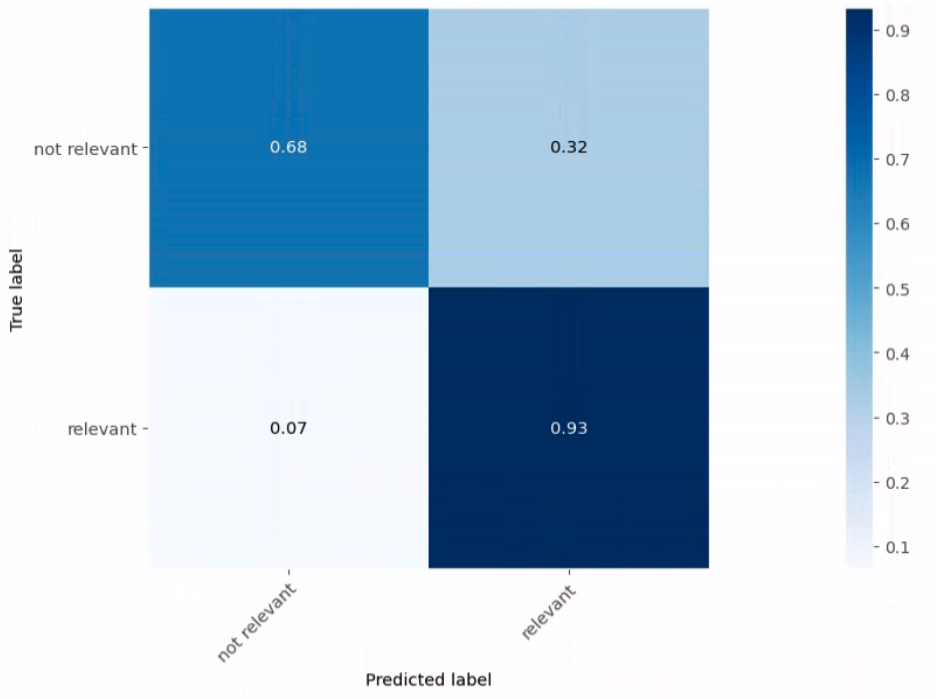


Figure 6.4. Normalised confusion matrix for the GPT-2 model.
This confusion matrix for the GPT-2 model shows the true positives (0.93), true negatives (0.68), false positives (0.32) and false negatives (0.07), in varying gradations of blue, with darker shades indicating a larger number, and lighter indicating smaller ones.

6.9.3 Anatomy Classifier

To facilitate the identification of body parts that might be mentioned in association with pain, a classifier was trained to run on sentences that have been labelled as containing relevant mentions of pain. The model was trained to classify these sentences into binary classes - “mentioned”, which indicates a body part is mentioned in relation to the pain, and “not mentioned”, which indicates a body part is not mentioned in relation to the pain. Similar to the previous classification task, 2 BERT-based models - BERT_base and SapBERT - were fine-tuned on gold standard annotations that indicated the presence or absence of body parts in relation to pain. In addition to these, a number of non-transformer-based models were trained, the 3 best performing of which are reported here: Stochastic Gradient Descent (SGD), Logistic Regression (LR) and Linear Support Vector Classifier (LSVC).

The gold standard annotations consisted of 4,028 sentences, with 63% of the sentences belonging to the “mentioned” class. Since this dataset is slightly imbalanced, similar to the other classifiers, cross-entropy loss was incorporated into the model's training. 3,222 sentences were used for training (further split into 2,899 for training and 323 for validation), and 806 sentences were retained as the test set.

The parameters used to train the model are outlined in Table 6.9, and the performance metrics are in Table 6.10.

Model	Tokeniser	Pre-processing	Other Parameters
SapBERT	cambridgeltl/SapBERT-from-PubMedBERT-fulltext	Tokenise Prepend sentence with special token [CLS] and append with	Epochs: 4 Batch size: 16 Optimizer: AdamW, learning rate 2e-5

		special token [SEP] Pad and truncate sentence to max length 105 (default is 511)	
BERT_base	bert_base_uncased	Tokenise Prepend sentence with special token [CLS] and append with special token [SEP] Pad and truncate sentence to max length 105 (default is 511)	Epochs: 3 Batch size: 16 Optimizer: AdamW, learning rate 3e-5
SGD LR LSVC	NLTK	Lowercase, stopword, white space and punctuation removal, lemmatise and tokenise	Tf-Idf vectorizer Default parameters from sklearn

Table 6.9. Model specifications

This table details the parameters for the training of each model, including any pre-processing measures taken, tokenisers used, and other training parameters such as epochs, optimizers and learning rates.

K-fold cross-validation was carried out for evaluation of the model, and 95% confidence intervals were calculated.

Model	Precision	Recall	F1-score (average from 10-fold cross-validation)

SapBERT	**0.94 (0.91-0.96)**	**0.94 (0.92-0.97)**	**0.94 (0.91-0.95)**
BERT_base	0.92 (0.89-0.95)	0.92 (0.90-0.95)	0.92 (0.90-0.93)
Best Performing non-BERT models			
SGD	0.86 (0.82-0.88)	0.86 (0.83-0.88)	0.86 (0.83-0.88)
LR	0.86 (0.84-0.89)	0.86 (0.84-0.89)	0.87 (0.83-0.89)
LSVC	0.86 (0.83-0.89)	0.86 (0.84-0.89)	0.86 (0.83-0.88)

Table 6.10. Evaluation Metrics (weighted average) on two BERT-based models and 3 non_BERT models, including 95% confidence intervals

As with the pain classifier, the SapBERT model performed the best at identifying sentences that mention anatomy. Error analysis conducted on the test set predictions using the SapBERT model indicated some common false negatives in instances that included abbreviations for “complained of” as co (“co back pain”, “co headache”), misclassification when special characters occurred before or after an instance of interest, positive mention of anatomy in relation to pain followed by negation for something else, such as “had headache declined painkillers”, “backpain not relieved by painkillers”, and when the body part is not immediately in the vicinity of the pain term like “pain in his lower back”. In addition, some common false positives included hypothetical mentions such as “if they develop sore throat”, and mentions within forms and questionnaires such as “do you feel chest pain no”.

CHAPTER 7: Knowledge Graph Embeddings

7.1 Foreword

As has been described in Section 1.1.2 of Chapter 1, an essential component of this project is to link pain entities within the clinical text to compositional structured knowledge of SNOMED CT modelled as a knowledge graph, to make use of the additional relations between the entities. Towards this aim, this chapter describes the development of KGE models that incorporate information from SNOMED CT and clinical text, and are used in a sentence classification task. For clarity, this chapter is split into 2 parts:

Part 1 (Section 7.2 - Section 7.8) titled "Development of a Knowledge Graph Embeddings Model for Pain" describes the building of KGE models. This work was done in collaboration with Dr Tao Wang. It was peer-reviewed and has been accepted at the AMIA 2023[18] conference. I contributed as the first author, designed the research, developed the KGE models, evaluated their performances, and drafted the manuscript. Dr Tao Wang reviewed the manuscript and provided suggestions for methodology improvement. The paper printed in the conference proceedings is available at:

Chaturvedi J, Wang T, Velupillai S, Stewart R, Roberts A. Development of a Knowledge Graph Embeddings Model for Pain. AMIA Annu Symp Proc. 2024 Jan 11;2023:299-308. PMID: 38222382; PMCID: PMC10785867.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10785867/

[18] https://amia.org/education-events/amia-2023-annual-symposium

The following sub-sections of this chapter reproduce the preprint of the paper, with some minor formatting adjustments to keep it in line with the thesis format. The content itself has not been altered.

Part 2 (Section 7.9) titled “Incorporating Knowledge into Classifier” details how the KGE models were incorporated into a Random Forest classifier and reports on the performance metrics of this classifier.

Development of a Knowledge Graph Embedding Model for Pain

Jaya Chaturvedi, MSc[1], Tao Wang, PhD[1], Sumithra Velupillai, PhD[1], Robert Stewart, MD[1,2], Angus Roberts, PhD[1,2]

[1]Institute of Psychiatry, Psychology and Neurosciences, King's College London, London, United Kingdom; [2]South London and Maudsley NHS Foundation Trust, London, United Kingdom

7.2 Abstract

Pain is a complex concept that can interconnect with other concepts such as a disorder that might cause pain, a medication that might relieve pain, and so on. To fully understand the context of pain experienced by either an individual or across a population, we may need to examine all concepts related to pain and the relationships between them. This is especially useful when modeling pain that has been recorded in electronic health records.

Knowledge graphs represent concepts and their relations by an interlinked network, enabling semantic and context-based reasoning in a computationally tractable form. These graphs can, however, be too large for efficient computation. Knowledge graph embeddings help to resolve this by representing the graphs in a low-dimensional vector space. These embeddings can then be used in various downstream tasks such as classification and link prediction.

The various relations associated with pain which are required to construct such a knowledge graph can be obtained from external medical knowledge bases such as SNOMED CT, a hierarchical systematic nomenclature of medical terms. A knowledge graph built in this way could be further enriched with real-world examples of pain and its relations extracted from electronic health records. This paper describes the construction of such knowledge graph embedding models of pain concepts, extracted from the unstructured text of mental health

electronic health records, combined with external knowledge created from relations described in SNOMED CT, and their evaluation on a subject-object link prediction task. The performance of the models was compared with other baseline models.

7.3 Introduction

Pain is a global health problem and is estimated to affect 1 in 5 adults worldwide (Goldberg & McGee, 2011). Pain is a massive burden on society in terms of costs related to medical care as well as loss of productivity (Rayner et al., 2016). A committee reviewing the public health significance of pain in the United States found that the total cost to society was greater than that estimated for heart disease, cancer or diabetes (Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education, 2011). People who experience chronic pain are more likely to develop emotional distress, which can create muscle tensions and increase pain. There is a known intersection between pain and mental health disorders, such as pain and depression (Bair et al., 2003), bipolar and psychotic disorders (Stubbs et al., 2015). Consequently, the impact of pain on mental health and quality of life is an active area of research. Pain is a common reason for people to seek medical attention (Gureje et al., 1998), and is therefore widely described in electronic health records (EHRs). In mental health EHRs, patients' experiences of pain are often recorded as free-text. EHRs have therefore become a valuable resource in the research of pain (Von Korff et al., 2020; Weng et al., 2020). There is substantial evidence to support an overlap between pain and mental health (Eisendrath, 1995; Viana et al., 2018). Compared to physical health conditions, more contextual information is generally required and therefore recorded about pain for patients with mental health conditions, making the clinical text within mental health EHRs a good source of such information.

Knowledge graphs (KGs) are large networks which allow for the representation of entities/concepts, along with their semantic types and relations to other entities as graphs (Ehrlinger & Wöß, 2016) . They have emerged as an efficient method of representing data as a heterogeneous graph, facilitating the visualization of and reasoning over complex data and its interconnected relationships, which further help reveal any hidden patterns and deduce new knowledge (Yoon et al., 2017). A KG typically consists of a set of fact triples, referred to as subject-predicate-object triples, or nodes and edges, or head-relation-tail triples (K M et al., 2018). For example, in a triple such as <paracetamol, relief, pain>, paracetamol is the subject/node/head, relief is the predicate/edge/relation, and pain is the object/node/tail. In this paper, we will use the terminology subject, predicate and object. KGs can be huge, making them impractical and computationally expensive to use. This issue is resolved by using KG embeddings, i.e., low dimensional representations in a vector space (Mikolov et al., 2013). Knowledge graph embeddings (KGEs) also assist in further enriching the data by representing the semantics of domain knowledge within the KGs (Chang et al., 2020). KGE models learn embeddings of the entities and relations based on scoring functions that predict the probability that a given triple is a fact, i.e. higher scores indicate a true triple or more likely to be factually correct. These scoring functions combine the embeddings of the triples using different intuitions. The two models used in this work are ComplEx (Trouillon et al., 2016) and TransE (Bordes et al., 2013), which are described in more detail in the Methods section. KGE methods have been applied to various biomedical use cases where data is linked to relevant ontologies and terminologies to predict relations (Alshahrani et al., 2017), understand gene to phenotype associations (Alshahrani & Hoehndorf, 2018), and predict disease comorbidity (Biswas et al., 2021). The multidimensional nature of pain (Merlin et al., 2014) makes it a good use case for application of such KGE methods. Other EHR-based use cases include

patient stratification and drug identification (Zou et al., 2022), and disease relation extraction (Lin et al., 2023).

This paper describes the development of KGE models of pain incorporating both pain concepts found within a mental health EHR database, and external knowledge about these concepts from a knowledge base, SNOMED CT (Stearns et al., 2001) (detailed in the Methods section), for use in research on the relationships between mental health, pain, and physical multimorbidities. Whilst it is common to build KGE models from knowledge bases, we have also incorporated information from the EHR, hypothesising that the addition of real-world language context will improve performance in downstream tasks on EHR text. Three models were constructed by varying the features that were included in the embeddings. The models were evaluated using a link prediction task, and comparisons made between these models. They were also compared with other biomedical and non-biomedical benchmark models that were publicly available. The best performing model will also be used in a text classification task, which is described in Section 7.9. Existing pain research is limited to the use of structured codes in combination with some clinical text from EHRs (Von Korff et al., 2020; Weng et al., 2020) or patient-focused questionnaires and interviews (Gureje et al., 1998; Rayner et al., 2016). There is limited research utilizing clinical text from within EHRs combined with external knowledge bases (Lin et al., 2023; Ye et al., 2023).

In summary, the task outlined in this paper involves a multi-step process. Initially, the data is prepared by forming triples, where pain terms from the lexicon (described in Chapter 3) are mapped to their parent and child nodes within a knowledge graph of SNOMED CT. These triples are then utilised to construct the first variation of KGE models. In addition, pain concepts identified through manual annotations within the CRIS dataset, as described in Chapter 5, are subjected to the same triple generation process by querying the SNOMED CT

knowledge graph. The resulting triples from both the lexicon and CRIS annotations are combined to build a second variation of KGE models. Furthermore, the third variation incorporates embeddings of sentences containing the pain concepts from the CRIS annotations, enriching the triples with contextual information. Each of these three variations include both TransE and ComplEx model architectures, resulting in a total of six distinct KGE models. These models are evaluated on a link prediction task. The best-performing model from this link prediction evaluation is then applied to a sentence classification task, as described in Section 7.9, leveraging its learned representations to classify clinical text as relevant to pain or not, using the same classes as described in Chapters 5 and 6. This systematic approach explores the potential of combining structured knowledge from a medical ontology with unstructured clinical text data, ultimately developing a robust model capable of accurately classifying pain mentions within clinical notes. To the best of our knowledge, this is the first time such KGE models have been developed for pain research. The models, and scripts used to develop them, are publicly available[19] and could be adapted for use in other areas of medicine.

7.4 Methods

7.4.1 Data Collection

EHR text was extracted from an anonymised version of a large mental health EHR from The South London and Maudsley NHS Foundation Trust (SLaM) through its Clinical Record Interactive Search (CRIS) data platform (Stewart et al., 2009). The infrastructure of CRIS has been described in detail with an overview of the cohort profile (Stewart et al., 2009). CRIS is

[19] https://github.com/jayachaturvedi/pain_kge_model

comprised of over 30 million documents and over 500,000 patient records (Stewart et al., 2009), averaging about 90 documents per patient (Velupillai et al., 2018).

SNOMED CT (Stearns et al., 2001) is one of the most commonly used medical knowledge bases in healthcare, and so has been used in various KGE models (K. Agarwal et al., 2019; Chang et al., 2020). The formal and hierarchical structure of SNOMED CT facilitates the classification of data into different taxonomic categories which combine various clinical concepts such as diseases and medications (Sastre et al., 2020). These add a level of semantics to clinical data by providing reference to different concepts and the relationships that exist between them, thereby enabling logical reasoning. In combination with natural language processing (NLP), such structured knowledge can help disambiguate concepts mentioned within the unstructured clinical notes of EHRs and produce more meaningful results (Stearns et al., 2001). One advantage of SNOMED CT over other biomedical terminologies is that it is designed as a compositional, post-coordinated system. This compositional design ensures that when SNOMED CT is used in different systems and contexts, it will still produce the same conceptual and computational meaning for concepts (Stearns et al., 2001). It could therefore be a valuable resource for NLP on clinical data.

As shown in Figure 7.1, a set of pain keywords were derived from a pain lexicon, development of which is described in Chaturvedi et al. (2021) and Chapter 3. A SQL query was run to extract free text documents from the CRIS database (no time or diagnosis filters were applied) that contained any of these pain keywords. Keyword searches can often lead to noise. To refine this, these documents were loaded onto a medical concept annotation tool, MedCAT (Kraljevic et al., 2021) which was used to pre-annotate all the pain concepts within the documents, linking each concept to a unique SNOMED CT ID (SCTID). Three medical student annotators manually reviewed these pre-annotations and marked them as relevant

mentions of physical pain or not, as described in Chapter 5. The annotation guidelines that were used by these students are publicly available[20]. In addition to this, SNOMED CT was used to generate the parent and child nodes of every term in the lexicon. This is described in Section 7.4.3 below. Since the sentences from CRIS were run through MedCAT for the annotation process, MedCAT provided SCTIDs for the various pain entities within the text. These SCTIDs were treated as subjects and used to look up and extract predicates and objects from the SNOMED CT KG.

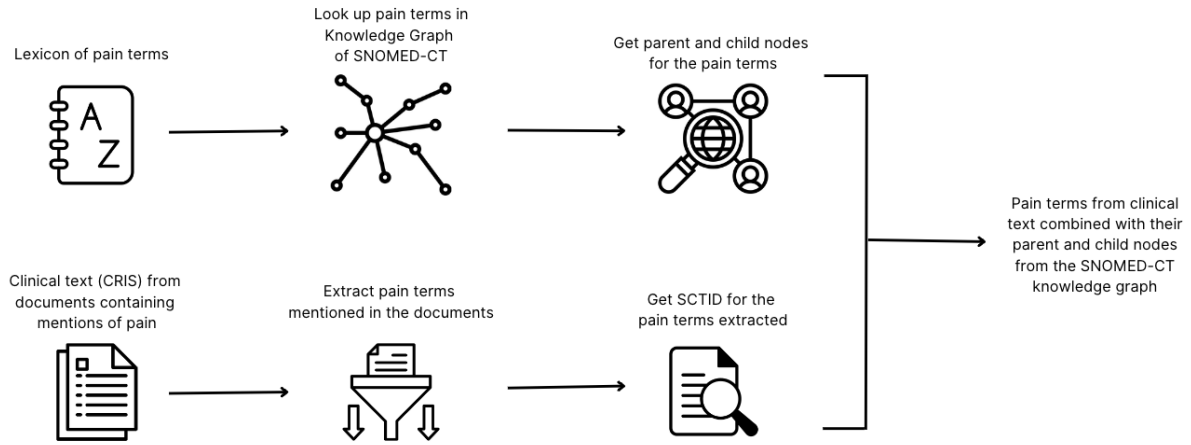


Figure 7.1. Creation of dataset for building KGE models
This diagram shows the process followed to create the dataset that was used to build the KGE models. Pain terms within the lexicon were looked up on the knowledge graph of SNOMED-CT to obtain their parent and child nodes, forming the triples that were used in one of the variations. In the second variation, these triples were combined with triples for the pain terms mentioned within the clinical text of CRIS. In the third variation, the sentences that contained the pain terms within CRIS were included as well.

7.4.2 Ethics and Data Access

The source clinical data are accessed through SLaM, the data custodian. Within a customised information governance framework, the Maudsley CRIS platform provides access to anonymised data derived from SLaM's electronic medical records. These data can only be

[20] https://github.com/jayachaturvedi/pain_in_mental_health/blob/main/Annotation%20Guidelines%20-%20Pain%20-%20for%20github.pdf

accessed by authorised individuals from within a secure firewall (data cannot be sent elsewhere)[21]. Ethical approval to use the data for research was granted by Oxford C Research Ethics Committee, reference 18/SC/0372.

7.4.3 Relation Extraction

A knowledge graph of SNOMED CT was developed using Clinical Knowledge Graph (CKG (Santos et al., 2022)), which was implemented in the Neo4J graph database format (Neo4j Inc., 2012). CKG contains 10 different ontologies, including SNOMED CT, and therefore contains every SNOMED CT concept and their parent/child nodes. A query written using Neo4j's Cypher query language was run on CKG to extract the first-order parent and child nodes for all the pain keywords derived from the lexicon. For example, the concept "abdominal pain" can have various relations as shown in Figure 7.2.

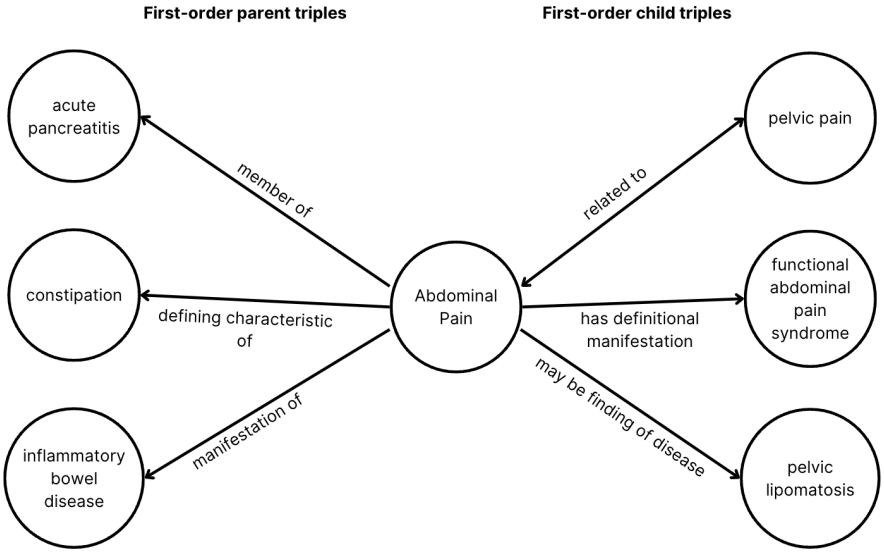


Figure 7.2. An example of first-order parent and child triples for the concept "abdominal pain"

This figure shows an example of the first order parent and child nodes for abdominal pain. The arrows denote whether the relation is uni- or bi-directional. The circles denote the nodes, and the arrows are the edges.

[21] Please contact cris.administrator@slam.nhs.uk for more information.

7.4.4 Knowledge Graph Embedding

Python version 3.7.16 and the AmpliGraph 1.4 library (Costabello et al., 2019) were used to develop KGE models using the triples generated from CKG. Two commonly used models are ComplEx (Trouillon et al., 2016), which uses tensor factorization (a three-way tensor is defined in the form of *n x n x m* where n is the number of entities (subject and object) and m is the number of relations (predicates) - the embeddings are calculated by factorizing this tensor), and TransE (Bordes et al., 2013), which relies on distance (the relationship between subject and object is interpreted as a translation vector so that the embedded entities connected by a relation have a short distance, i.e., distance-based functions in the Euclidean space). ComplEx was used since it is considered better at representing multi-dimensional data and preserving asymmetry between concepts such as those defined in biomedical ontologies (Alshahrani et al., 2021). This was compared to TransE (Bordes et al., 2013) which is commonly used as a benchmark. These two models were chosen because each of them have strengths that may be advantageous in the EHR setting. TransE models asymmetry, inversion and composition, the latter being most useful for SNOMED CT which is inherently compositional in nature. However, TransE lacks the ability to model symmetry, and one-to-many relations. ComplEx models symmetry, asymmetry, inversion, and one-to-many relations. However, it lacks the capacity to capture composition (Sun et al., 2019).

Three variations of the KGE models were constructed, each variation included both ComplEx and TransE:

Variation 1: The triples of the pain keywords from the lexicon were used in the development of these models. This variation does not include any EHR data.

Variation 2: The triples of pain keywords from the lexicon were combined with the pain concepts and their triples within the sentences from CRIS data to form the final dataset for

development of these models. This was to ensure incorporation of pain concepts mentioned within CRIS data, but either not in the lexicon, or in the lexicon as variants. For example, “response to pain” and “on examination - painful ear” are concepts found within the sentences, but referred to as “pain” and “ear pain” within the lexicon.

Variation 3: In addition to the data used in variation 2, the embeddings of the sentences that contained the pain concepts within CRIS were included in the development of these models. This was to capture the context of the sentences that contained the pain concepts, and makes use of the fact that both models, ComplEx and TransE, represent the triples, pain concepts, and sentences in a shared continuous vector space. This allows for meaningful calculations and enables the models to be used for tasks like link prediction efficiently. In order to represent sentences in the same embedding space, averaging or pooling of the embeddings of the individual words is undertaken, which results in a single vector representing the entire sentence. This shared embedding space between the triples, pain concepts and sentences helps capture the semantic relationships between them.

7.4.5 Link prediction

The data for each variation was randomly split into training and test sets in the proportion of 80:20. The training set was used to build the models, and the test sets were used for evaluation of the models on a link prediction task i.e. the model is given the subject-predicate and asked to predict the object, and vice versa. The link prediction task is conducted purely for evaluation of performance of the models. Default Ampligraph parameters were used (listed in Table 7.1). Since the data used for the 3 variations are inherently different as they include sentence embeddings while the other variations do not, it was not feasible to use the same data for training all the variations, although the triples from variation 1 are common to all 3 variations. To ensure fair comparisons of these variations, a consistent method was used for

splitting the date into 80:20 proportions for each variation. 10-fold cross validation was also conducted on the 3 variations, to compensate for the differences in the datasets and make the evaluation more robust.

Model	Parameters
ComplEx	[1]batches_count: 100 seed: 555 epochs: 10 [2]k: 150 [3]eta: 10 loss: 'multiclass_nll'
TransE	[1]batches_count: 100 seed: 555 epochs: 10 [2]k: 150 [3]eta: 10 loss: 'pairwise'

Table 7.1. KGE model parameters

[1]batches_count: number of batches in which the training set is split during training
[2]k: dimensionality of the embedding space
[3]eta: number of negative, or false triples, generated for each positive, or true triple, during training
Default loss functions for each model were used

For the link prediction task, the metrics will be reported on Mean Reciprocal Rank (MRR) and Hits@N, which are two popular metrics for this type of evaluation task. The ranks in MRR indicate the rank at which the test set triple was found when performing link prediction using the models. Mean Reciprocal Rank (MRR) is a measure of how well a KGE model can predict the missing link (either the subject, the object, or the relation) based on the embeddings learned by the model. For example, given a subject (Ex: "abdominal pain"), and relation (Ex: "may be finding of disease"), the KGE model predicts the rank of each possible object. The reciprocal rank is then calculated for the correct object. In our example, if the model ranked

the correct answer (Ex: “pelvic lipomatosis”) as 1st choice, the reciprocal rank would be 1, whereas if it were ranked as 2nd choice, the reciprocal rank would be 0.5. A higher MRR indicates that the model is more accurate at finding the correct relationship. Hits@N computes how many elements of a vector of rankings was in the top n positions. Hits@N measures the percentage of correctly predicted entities in the top N ranked results. Hits@1 measures the percentage of correctly predicted entities in the top 1 ranked result, and Hits@10 measures the percentage of correctly predicted entities in the top 10 ranked results. The higher the Hits@N score, the better the model is at predicting missing links in the knowledge graph. During link prediction, the original triples in the knowledge graph are corrupted to form negative examples. For example, for a given positive triple (head, relation, tail) in the knowledge graph, negative examples are created by replacing either the head or the tail entity with some randomly chosen entities. These corrupted triples serve as distractors and are used to evaluate the model's performance. For link predictions, the metric of choice is generally Hits@N since it focuses on the model’s ability to rank the correct entity within the top N positions. MRR emphasises the overall performance while Hits@N emphasises more on results in top ranking However, it is better to consider multiple metrics, such as MRR in combination with Hits@N, so that we can get a comprehensive understanding of the model’s performance.

The two models developed here (ComplEx and TransE) were compared to biomedical (trained on SNOMED CT (Chang et al., 2020)) and non-biomedical benchmarks (trained on FreeBase (AmpliGraph, 2019a)) that are commonly used for such tasks.

7.4.6 Pipeline for use

The KGE models were also used in combination with a classifier, which is detailed in Section 7.9. When incorporating the KGE models with a classifier, a few additional steps are required,

and a pipeline for how these models can be used to classify text and predict labels for new data has been detailed below.

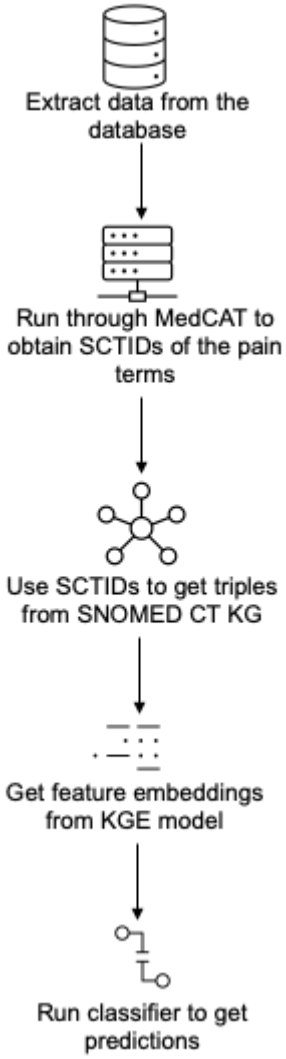


Figure 7.3. Pipeline for classification incorporating KGE

This flow diagram shows the pipeline that has to be followed in order to get classifier predictions on new data. Once the data is extracted from the database, it has to be run through MedCAT in order to obtain the SCTIDs for any pain terms. These SCTIDs are then used to obtain the parent and child nodes for the pain terms from the knowledge graph of SNOMED CT. Once we have the triples for each pain term, their feature embeddings are extracted from the trained KGE model, and then these feature embeddings are used to get predictions of classes from the trained classifier.

As shown in Figure 7.3, the embeddings from the previously trained KGE model will be used to classify instances using the random forest classifier. This process is more time-intensive

than the other classifier models that do not use the KGE model because it incorporates an additional step of obtaining SCTIDs to extract triples for each pain entity mentioned in the text. This process also increases the size of the dataset being fed to the classifier for predictions (despite the underlying number of documents being the same in both instances of random forest with and without the KGE model), so the runtime is significantly longer when dealing with a large number of sentences.

7.5 Results

7.5.1 Data Statistics

The pain concepts from the lexicon were used to generate triples from the SNOMED CT knowledge graph within CKG (total of 15,336 triples). A portion of these (training set: 80%) were used to generate the KGE models in variation 1. 5,644 sentences from the CRIS data were identified as containing a total of 206 unique pain concepts. These were merged (triples from CKG and the 206 pain concepts from CRIS) to form the final dataset for building the KGE models, 80% of which was used in building the models in variation 2. 80% of the 5,644 sentence embeddings were included in the dataset used to build the models in variation 3. The data sources, as seen from the details in the table below, are not entirely comparable due to addition of extra information in each variation.

Data Source	Number of pain terms	Number of triples from SNOMED CT (using CKG)	Included in Variation
Pain Lexicon	382	15,336	1, 2 and 3
CRIS – from the gold standard annotations	206	25,520	2 and 3

CRIS – embeddings from sentences that contain the pain terms from the gold standard annotations	206	51,040	3

Table 7.2. Data table detailing the different data sources, number of triples, and variations involved.

The first source includes terms from the pain lexicon only. The second source contains the pain terms from the gold standard clinical text annotations within CRIS. The third source adds on embeddings from sentences that contain the pain terms along with the pain terms (this is why the number of pain terms is the same for the second and third source – the addition is the sentence embeddings added to the triples as <pain term>-<contains context>-<sentence embedding>)

The most frequent triple in our data was "*pain"-"may be treated by"-"aspirin".* The top 5 subjects and predicates are listed in Table 7.3.

Top 5	From pain lexicon		From CRIS data	
Subject	Pain	15%	Pain	42%
	Headache	6%	Chest pain	6%
	Abdominal pain	6%	Abdominal pain	4%
	Rheumatoid Arthritis	5%	Headache	4%
	Spasm	4%	Sore sensation quality	3%
Predicate	inverse is a	25%	may be treated by	36%
	may be treated by	23%	may be finding of disease	18%
	may be finding of disease	10%	may be prevented by	16%
	classifies	6%	inverse is a	8%
	may be prevented by	5%	classifies	2%

Table 7.3. Top 5 subjects and predicates (Objects not included because the frequency of each was very small (<1%))

7.5.2 Results of link prediction

The performance metrics for the link prediction task of the KGE models are given in Table 7.4. The variation that included pain concepts as well as sentence embeddings from the EHR data (variation 3) performed best, and overall the ComplEx model performed better than TransE in all instances, with an MRR of 0.88.

Models		Performance Metrics		
		MRR	**Hits@10**	**Hits@1**
Non-biomedical benchmark	**ComplEx**	0.32	0.50	0.35
	TransE	**0.31**	**0.50**	**0.35**
Biomedical benchmark	**ComplEx**	0.46	0.65	0.36
	TransE	0.34	0.59	0.21
Pain KGE without EHR data (Variation 1)	**ComplEx**	0.15	0.27	0.11
	TransE	0.18	0.33	0.12
Pain KGE with pain concepts from EHR data (Variation 2)	**ComplEx**	0.79	0.86	0.74
	TransE	0.30	0.48	0.20
Pain KGE with pain concepts and sentence embeddings from EHR data (Variation 3)	**ComplEx**	**0.83**	**0.87**	**0.80**
	TransE	0.29	0.41	0.23

Table 7.4. Performance metrics of the two models (ComplEx and TransE) for the three variations, compared to biomedical benchmarks that were trained on SNOMED CT (Chang et al., 2020) and non-biomedical benchmarks trained on FreeBase (AmpliGraph, 2019a)

7.6 Discussion

This paper describes the development of KGE models with and without utilizing EHR data about pain from a mental health EHR database, combined with external knowledge from SNOMED CT. A link prediction task was used to evaluate the performance of the different models and variations, and will not be used in any clinical tasks within this project. The metrics used to evaluate the link prediction, MRR and Hits@N, provide insights into the model's ability to rank true relationships higher than false ones, thereby indicating that the model has learnt the required embeddings that would capture the semantic relationships between the entities. This would in turn be useful and transferable to classification tasks. However, additional evaluations will be carried out on the classification tasks as well, by utilising the relevant metrics for classification such as precision, recall and F1-score, as detailed in Chapter 8. The ComplEx model performed better than TransE in most variations. This could be because ComplEx has the ability to capture nuanced relationships between entities and relations by representing them as complex vectors. TransE, on the other hand, uses simple vectors. ComplEx performed best in variation 3, which incorporated sentences from the EHR in addition to the pain concepts. Incorporation of more data into the construction of the KGE model meant more relationships between entities, especially of the one-to-many nature, which is a strength for ComplEx-based models. They are able to model multiple relations between entities, while TransE was designed to handle one-to-one relations. The addition of sentence embeddings into the models from variation 3 would also have meant more features to learn from, and therefore better performance. Another strength for ComplEx is its ability to handle noisy data, which EHR data is renowned for. The use of simple vectors in TransE means it is impacted by noise in data. TransE performed best on the non-biomedical benchmark trained on FreeBase. This could be because such data is not as noisy as EHR

data. Overall, these two models performed better on link prediction when compared to biomedical and non-biomedical benchmarks. The dataset used for generation of the KGE models in this work is quite small, and specific to pain concepts, which could be why it performed better than the larger biomedical and non-biomedical benchmarks.

7.7 Conclusions

The ambiguous nature of pain and the complexity of how it is described within text highlights the need for additional information from external knowledge bases to supplement the data available with EHRs, in combination with contextual information from the sentences that contain information about pain. While recent literature has also incorporated such contextual information in their work, they lack the advantage of leveraging the compositional nature of a knowledge base such as SNOMED CT, and instead rely on sources such as ICD-9 or ICD-10 (Lin et al., 2021), DrugBank (Sastre et al., 2020) or niche databases such as the traditional Chinese medicine knowledge base (Ye et al., 2023).

As part of future work, the ComplEx KGE model built in variation 3 will be used in a downstream binary sentence classification task, to classify sentences as relevant to pain or not. Description of pain in mental health records is mostly restricted to the unstructured free-text of the records. By developing a method to extract information about pain from text, we are able to use that information in studies of pain in the context of mental health. It will allow for better understanding of whether patients with certain mental health disorders report more pain. This could potentially help in early detection of such pain, thereby improving patient outcomes that could have deteriorated due to long durations of untreated pain symptoms (Husain & Chalder, 2021). Output from this will then be used to explore associations between pain and mental health, and comorbid pain as a predictor of adverse outcomes for people

with mental disorders. We have shown that the KGE models that combined information from structured knowledge and real-world textual data from EHRs performed best, which shows potential in performing better at downstream tasks that classically only use EHR data. These results will be compared to those of classifiers built for the same use-case without the incorporation of any external knowledge. While the pipeline to use the classifiers in combination with the KGE model will be more complex due to the need for added information, such as triples for all the pain concepts mentioned within the clinical notes, the benefit of better performance will ensure more accurate classification, and therefore better quality of pain information extraction, which will in turn feed back into better pain research for patient care and pain management. All code used to generate the triples using CKG, the Python code for building and evaluation of the KGE models, and the models themselves, are openly available on GitHub[22]. The exception is variation 3, whose model has not been made available since it contains sentences from the CRIS database.

7.8 Funding and Acknowledgements

AR was part-funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. RS and AR are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RS is additionally part-funded by the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust, and by the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386). AR and

[22] https://github.com/jayachaturvedi/pain_kge_model

RS were additionally part-funded by the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities. JC was supported by the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. TW was supported by the Maudsley Charity and an Early Career Research Award from IoPPN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7.9 Incorporating Knowledge into Classifier

7.9.1 Introduction

This section describes the incorporation of the ComplEx model (detailed in Section 7.5 and Section 2.3) into a random forest classifier model. The objective was to incorporate domain knowledge from the KGE model into the classifier. Since the KGE model will have effectively captured the relationships and associations between entities within the clinical text and the SNOMED CT KG, these valuable semantic relationships have the potential to enhance the classification process by leveraging the domain-specific information at its disposal and creating better feature representation, thereby improving the robustness of the classification model.

7.9.2 Building the Classifier

The gold standard annotations used to develop the classifier applications described in Section 5.5.1 of Chapter 6 have been used to build this classifier. As described in Section 7.4, the SCTIDs obtained for the various pain entities within the text through MedCAT during the annotation process were used to extract predicates and objects from the SNOMED CT KG. This KG was built using CKG and has been described in more detail in Section 7.4 of this chapter. The final gold standard annotations now consist of the patient and document identifiers, the text for classification, and the value within the text i.e. the pain entity or subject, its predicates and its objects. This data is then split into train and test sets for the development of the classifier. In order to prepare the data for training the classifier model, we obtain the features (knowledge embeddings) for the train and test data from the ComplEx KGE model. However, since each pain entity can have multiple predicates and objects, ensuring that the

embeddings of all the predicates and objects of that pain entity were aggregated before being used for training was essential.

Like the classifier mentioned in Section 6.9.1 of Chapter 6, sklearn's Random Forest classifier model was used with default parameters. The model was trained, and performance metrics were obtained from validation on the test set.

7.9.3 Classifier Performance

The objective of the classifier was to classify sentences as relevant/ not relevant, similar to the task described in Chapter 6. The table below shows the performance metrics for the random forest classifier built by incorporating embeddings from the KGE model.

Model	Precision	Recall	F1 score
Random Forest with KGE model	0.96 (0.92-0.99)	0.93 (0.90-0.96)	0.94 (0.90-0.97)

Table 7.5. Performance Metrics for the Random Forest model incorporating the ComplEx KGE model (variation 3), including 95% confidence intervals

Error analysis on the test set showed false positives on some hypothetical mentions, such as "if pain develops", and pain that does not exist, such as "interpreting other sensations as pain" and "pain hallucinations". The false negatives were instances such as "history of chest pain" and mentions of existing pain getting worse, such as "if pain increases" and "if worsening abdominal pain".

7.9.4 Conclusion

While the process for incorporating the KGE model into a classification task is more computationally and time intensive, this work shows the potential for such incorporation of external knowledge into an NLP task. The SapBERT model (described in Section 6.3 of

Chapter 6) follows a similar principle and is much easier to implement at scale, so it was used for the prevalence study in Chapter 9.

CHAPTER 8: Comparison of Outputs

The preceding chapters have described the development of two categories of classifier models for the identification of pain information from clinical notes: ones that incorporate external domain knowledge and ones that do not. The task was to classify sentences as relevant/not relevant where relevant refers to sentences that include mentions of physical pain affecting the patient, and not relevant refers to sentences that include negated, hypothetical, metaphorical or no mentions of physical pain affecting the patient. This chapter compares four classifier models, two from each category, and assesses the impact of incorporating structured biomedical knowledge into such classifiers. The models include two conventional approaches, BERT_base and Random Forest (described in Section 6.3 and Section 6.9.1 of Chapter 6, respectively), as well as two models augmented with domain knowledge - SapBERT, which incorporates UMLS and Random Forest with KGE, which incorporates knowledge from SNOMED CT (the former is described in Section 6.3 of Chapter 6, and the latter in Section 7.9 of Chapter 7). The performance of these models is evaluated using precision, recall and F1 score performance metrics. Additionally, the models are run on sentences from a cohort of patients (a description of this cohort is detailed in Section 9.4 of Chapter 9) within CRIS. The frequency of each label in the output of the classifiers was compared between the four models. The objective of this comparison was to give a measure of differences and overlaps in class assignments by each model. Beyond aggregate performance metrics, comparing the frequency of predicted labels between the models will help understand whether similar performance metrics mean similar class assignments, and if not, then what factors could affect this disparity. This is important to understand because it will help inform decisions on appropriate model selection, and the implications of using different types of models with and without knowledge. The results of this chapter will shed

light on the potential advantages and limitations of knowledge-augmented models for the classification of sentences within the clinical text with respect to pain.

8.1 Comparison of Classification Metrics

The gold standard annotations described in Section 5.5 of Chapter 5 were used as training and testing data, in the proportion of 80/20, for developing the four classifier models listed below. The performance metrics (with 95% confidence intervals) reported below were calculated based on the predictions made on the testing data. These are summarised in Table 8.1 and Figure 8.1 below. The SapBERT model, which is pre-trained on UMLS and so incorporates external domain knowledge, performed the best overall, and the incorporation of KGE within the Random Forest model brought its performance close to that of the transformer-based models.

Model	Precision	Recall	F1-score
BERT_base	0.96 (0.94-0.97)	0.98 (0.97-0.99)	0.97 (0.96-0.98)
SapBERT	**0.98 (0.97-0.99)**	**0.99 (0.98-0.99)**	**0.98 (0.98-0.99)**
Random Forest	0.86 (0.84-0.88)	0.86 (0.83-0.88)	0.85 (0.83-0.87)
Random Forest with KGE	0.96 (0.94-0.98)	0.93 (0.90-0.95)	0.94 (0.92-0.96)

Table 8.1. Evaluation metrics for the 4 models, including 95% confidence intervals in brackets.

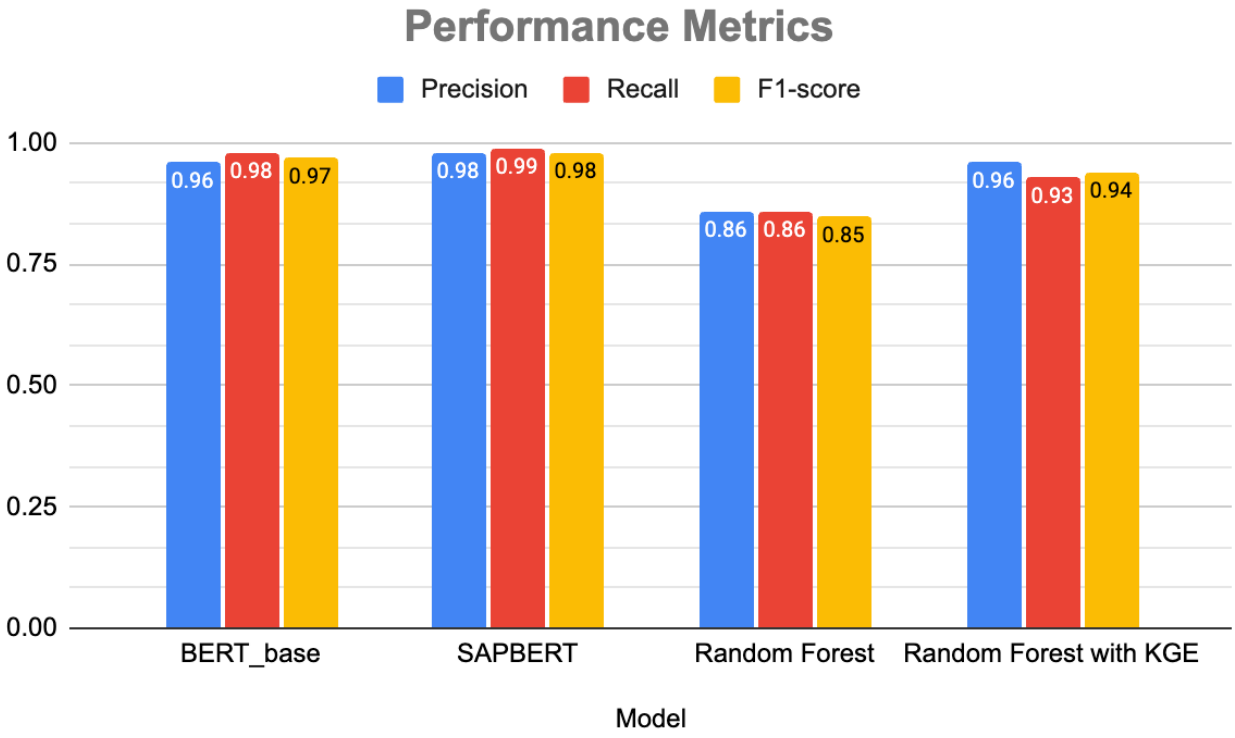


Figure 8.1. Performance Metrics for the 4 classifier models, showing precision, recall and F1-scores.

8.2 Comparison of Predicted Labels

Any overlaps between the predictions were examined by calculating the inter-model agreements (equivalent to inter-annotator agreement calculations). The output frequencies of the classifiers on the cohort have been outlined in Table 8.2. Here, Class 0 refers to "not relevant" and class 1 refers to "relevant" as described in Chapters 5 and 6. The original split of the gold standard annotations was 72% Relevant/Class 1 and 28% Not Relevant/Class 0.

Model	Class 0	Class 1
BERT_base	14%	86%
SapBERT	22%	78%

Random Forest	13%	87%
Random Forest with KGE	23%	77%

Table 8.2. Comparison of predicted labels between the 4 models

Class 0 refers to not relevant to physical pain and Class 1 refers to relevant to physical pain.

When compared to BERT_base, SapBERT was better able to deal with negation and false positives, such as a person's name, and was more frequently correct at classifying these instances (92% of the time when manually checked on a sample of 1,200 sentences). However, these instances are not very common in the dataset, accounting for only 1% of the sample of 1,200 sentences, and the model correctly classified it 11 out of 12 times.

Table 8.3 displays the results of inter-model agreement calculations between the four classifiers. These metrics quantify the level of agreements between model outputs, similar to inter-annotator agreements for human annotators.

To determine the inter-model agreement, F1 scores were calculated by treating one model's predictions as the "ground truth" and the others as the "predictions". Instances classified as pain (class 1) by both models were considered true positives. Instances classified as pain only by the first model were considered false negatives, and those classified as pain only by the second model were false positives. Precision and recall were computed from these values and used to derive the F1 score, as shown below:

Precision = True Positives / (True Positives + False Positives)

Recall = True Positives / (True Positives + False Negatives)

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

This process was repeated with each model in turn acting as the reference or "ground truth", yielding an F1 score for each model pair. The average F1 provided an overall agreement measure:

Average F1 = (F1-score for model 1 + F1-score for model 2) / 2

Higher average F1-scores indicate greater consensus between model outputs. Similarly, higher Cohen's kappa values reflect better alignment of predictions. Higher scores on both metrics suggest the models produce consistent classifications given the same input data.

Models	Inter-model F1-score	Cohen's k
Random Forest with KGE and SapBERT (both KGE models)	0.83	0.62
Random Forest and Random forest with KGE	0.93	0.85
Random Forest and BERT_base (both non-KGE models)	0.93	0.84
SapBERT and BERT_base (both BERT models)	0.91	0.80

Table 8.3. Inter-model Agreements – F1 score and Cohen's k

This table shows comparisons of inter-model agreements between various combinations of the 4 models. This comparison is done using the F1 score and Cohen's Kappa metrics.

The overlap of class 1 predictions between the different models is demonstrated in the Venn diagrams in Figure 8.2. These were generated using a Python library called Matplotlib-venn[23], version 0.11.9.

[23] https://pypi.org/project/matplotlib-venn/

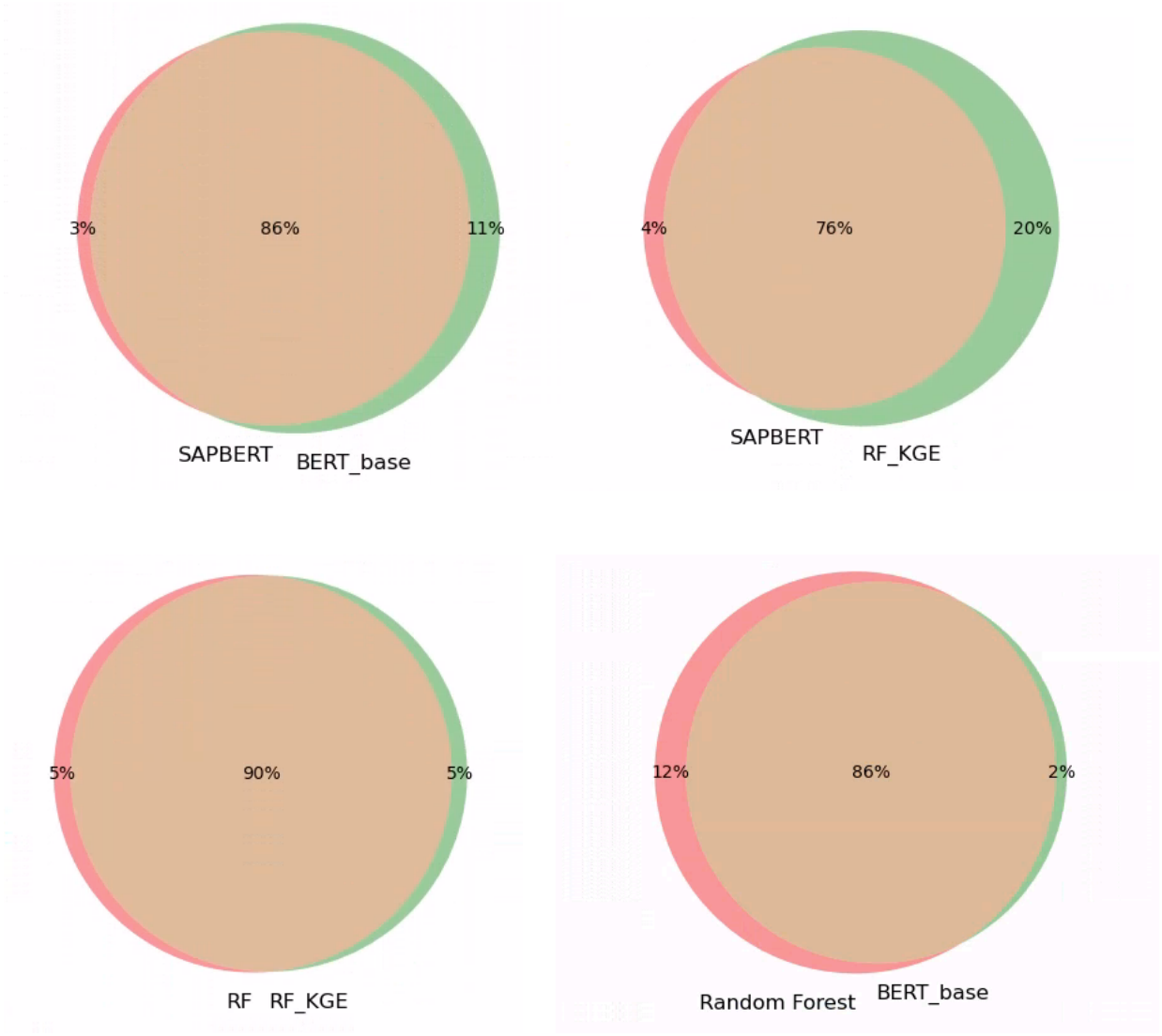


Figure 8.2. Overlap of predictions of class 1 between the four classifiers
These Venn diagrams show how the class 1 predictions overlap between the four different classifier models. The Venn diagrams are kept proportional to their overlap for a better visual representation.

8.3 Discussion

The comparative analysis described in this chapter reveals SapBERT as the top-performing model for extracting pain information from clinical notes, with an F1-score of 0.98. The conventional BERT_base model achieved a close second at 0.97. The superiority of these transformer-based models likely stems from their ability to incorporate sentence context through self-attention mechanisms (as described in Section 2.2 of Chapter 2). By contrast,

the non-neural Random Forest models performed worse, potentially struggling with the nuanced semantic features. Additionally, the models that incorporated external domain knowledge, SapBERT and Random Forest with KGE, demonstrated similar proportions of the frequencies of labels, with approximately 78% of sentences being labelled as class 1 versus 87% by the BERT_base and Random Forest models. The distribution of the labels by the knowledge-augmented models is more similar to those noticed in the gold standard annotations (described in Section 5.5.2) where 72% of the annotations were relevant, indicating that the models without knowledge are possibly generating more false positives. While this difference in proportions between the two categories of models is minor, it could suggest better detection and differentiation between the classes by the models using background knowledge. Also, as seen in Table 8.1, the BERT based models showed higher recall compared to precision, which is ideal for this particular use case. Higher recall indicates fewer missed relevant pain sentences, thereby minimising false negatives. With the downstream objective of identifying relevant mentions of pain, failure to capture all relevant mentions within the text could lead to exclusion of patients who were talking about pain thereby underrepresenting this population. Future work should consider optimising recall by reducing false negatives through a rigorous error analysis.

The random forest variants particularly highlight the impact of external knowledge, as the base algorithm was identical in both. Integrating domain knowledge through the KGEs appears to enhance the performance overall. However, the random forest models lack the language understanding capabilities that are inherent in the transformer models. This was apparent with hypothetical mentions, where both random forest models misclassified mentions such as “life is too painful”, “fear of pain”, and “continue to ask about pain” more frequently than the transformer-based models. Despite this, grounding the model in structured domain knowledge appears to improve performance in identifying relevant pain information

within clinical notes. These results demonstrate that while incorporating external knowledge provides benefits, the underlying model architecture remains crucial, as models like random forest, without contextual information, will still struggle with nuanced language. Therefore, while domain knowledge helps performance, it does not override the need for incorporating contextual information to effectively perform classification tasks.

While the frequencies of the predicted labels were similar, the overlap between SapBERT and random forest with KGE was the lowest compared to others and showed the lowest inter-model agreement, indicating differences in classification despite similar performance metrics and class proportions. They achieved 76% overlap on class 1 labels, with an F1-score of 0.83 and a Cohen's kappa of 0.62. This could be attributed to the differences in the knowledge sources for these two models. SapBERT incorporates UMLS containing 4M+ concepts and 10M+ synonyms (as discussed in Section 2.2 of Chapter 2), while the pain KGE model uses a subset of 15,336 SNOMED CT concepts (as described in Section 7.4 of Chapter 7). Additionally, SapBERT's transformer architecture is better at capturing semantic context when compared to the ComplEx KGE model's tensor factorisation-based embeddings. The breadth of knowledge and the advantages of contextual modelling within the SapBERT model potentially explain its improved performance at identifying pain mentions within clinical notes. With over 10 times more concepts and synonyms from UMLS incorporated into the model during its pre-training, SapBERT is better equipped at identifying more nuanced pain mentions that might be missed by the model incorporating KGEs from a smaller subset of domain knowledge. While both models that incorporate knowledge have shown higher classification performance, their discrepancies highlight the impact of knowledge sources and integration designs on the models' behaviours in classification tasks on clinical text.

When examining the differences in performance between the two transformer-based models, the differences in tokenisation used by SapBERT and BERT_base are likely a contributing factor affecting their performances. While both models utilise WordPiece tokenisation methods (described in Section 2.2 of Chapter 2), BERT_base split words like “ibuprofen” into ib, ##up, ##ro, ##fen, while SapBERT was able to maintain the entire word without any splitting, due to it being pretrained on biomedical concepts. This difference in keeping medical terms unified could allow SapBERT to better recognise these important keywords during classification. Splitting terms may hinder BERT_base's ability to directly match full semantic concepts. For example, "sciatic" becomes "sci", "##atic", and “migraine” was split into mig, ##raine in BERT_base, but both terms were maintained as single tokens in SapBERT. With false mentions like painting and painter (picked up because of the “%pain%” keyword), SapBERT splits the terms into pain, ##ting and #pain, ##ter, respectively. At the same time, BERT_base uses the whole word 'painting' or 'painter'; despite this, BERT_base classifies it incorrectly. It could be that SapBERT looks at what follows 'pain' to determine if it should be classified as class 1 or 0. Bearing all these differences, SapBERT's medically informed tokenization likely enables more accurate concept matching, keyword isolation, and contextual reasoning compared to BERT_base's generic WordPiece approach. The ability to recognise complete medical terms while breaking apart false matches potentially provides SapBERT with better performance on ambiguous pain mentions within the clinical text.

As can be observed from these results, it is possible for classifiers to achieve the same overall performance scores and yet make distinct errors on individual examples, where certain models might be more prone to have false positives or be more conservative and produce more false negatives. This discrepancy in labelling implies that the different models potentially rely on different features. This highlights that comparing predictions instance by instance can

unveil nuances that cannot be perceived from scores alone. This guides research into refining model architecture and features for optimal application-specific performance.

Additionally, a baseline using keyword search alone suggested 67% of instances containing pain terms compared to 33% without explicit keywords. However, this breakdown is not representative of truth, as only 52% of the sentences that contained pain related keywords were relevant after applying the classification models on them. By distinguishing sentences with spuriously matching terms from relevant pain mentions, our approach significantly improves upon this misleading baseline distribution. Simply querying lexicon matches would substantially overestimate the cohort's relevant pain mentions.

With each classifier model showing unique precision and recall trade-offs, assembling an ensemble pipeline combining the strengths of the different models could promote more balanced performances. A combination of transformer-based architecture with KGE models could yield further improvements. Beyond this, the validated models constructed in this project hold tremendous value for unlocking a variety of epidemiological investigations and clinical applications. Researchers can implement these models to accurately extract cohorts and pain mentions across massive corpora of mental health notes and derive prevalence rates associated with various diagnoses. Additionally, the classifiers could be integrated into medical record systems to trigger automated prompts for pain screening of high-risk subgroups. More broadly, the methods open the door to explorations of recorded pain patterns in other databases and settings as well.

8.4 Conclusion

This chapter compared four classifiers - two incorporating external knowledge and two without external knowledge. Any overlap between the predictions made by these models was

examined, and performance metrics were compared. The two BERT-based models performed better than the Random Forest models and showed considerable overlap in their predictions. Similarly, the two Random Forest models showed overlap as well. The SapBERT model performed the best overall and was easier to implement than its Random Forest counterpart. This ease of implementation can be attributed to its availability within the HuggingFace repository. The Random Forest with KGE model was more computationally intensive and took longer while training and running on new unseen sentences for predictions. Regardless of these differences, the SapBERT and Random Forest KGE models performed better than their counterparts, BERT_base and Random Forest without KGE, although the difference between the two transformer-based models was marginal. In addition, the incorporation of external knowledge within the Random Forest model brought its performance close to that of the transformer-based models. This improvement in performance shows that, as hypothesised in Section 1.1.1 of Chapter 1, linking entities from clinical text to structured knowledge modelled as knowledge graphs does improve the performance of classification tasks and shows potential to improve EHR analysis for pain and mental health research. These results also highlighted the importance of not relying on performance metrics alone but also considering deeper investigations into the individual labels classified and error analysis, as this can indicate individual differences in decisions being made by the different models.

CHAPTER 9: Distributions of Recorded Pain

9.1 Foreword

As mentioned in the previous chapter, upon comparison of all the classifier models, SapBERT performed the best. For this reason, SapBERT was run on documents from a cohort of patients within CRIS, for the purposes of identifying patients whose documents had relevant mentions of pain. In addition to this, the sentences labelled by SapBERT as relevant for pain were also run through the SapBERT anatomy classifier (described in Section 6.9 of Chapter 6). This chapter describes the data extraction criteria for this cohort, the application of the classifier to the sentences within the cohort, and the use of the generated meta-data to carry out a study of the distribution of recorded pain by sociodemographic and clinical status in mental health service users.

This work has been accepted at the journal *BMJ Open* for publication (ID: bmjopen-2023-079923) and was done in collaboration with Mark Ashworth. I contributed as the first author, designed the research, analysed distributions of the pain and anatomy mentions, and drafted the manuscript. Mark Ashworth provided feedback on the LDN data and its implications. All authors provided inputs on the manuscript.

The following sub-sections of this chapter reproduce the submitted paper, with some minor formatting adjustments to keep it in line with the thesis format. The content itself has not been altered.

Distributions of Recorded Pain in Mental Health Records: A Natural Language Processing Based Study

Jaya Chaturvedi[1*], Robert Stewart[1,2], Mark Ashworth[1], Angus Roberts[1]
[1]Institute of Psychiatry, Psychology and Neurosciences, King's College London
[2]South London and Maudsley NHS Foundation Trust
*Corresponding author: jaya.1.chaturvedi@kcl.ac.uk

9.2 Abstract

Objective

The objective of this study is to determine demographic and diagnostic distributions of physical pain recorded in the clinical notes of a mental health electronic health records database by utilising natural language processing and to examine the level of overlap in recorded physical pain between primary and secondary care.

Design, Setting and Participants

The data were extracted from an anonymised version of the electronic health records from a large mental community and secondary healthcare provider serving a catchment of 1.3M residents in south London. These included patients under active referral and aged 18+ at the index date of July 1, 2018, and had at least one clinical document (>=30 characters) associated with their record between July 1, 2017 and July 1, 2019. This cohort was compared to linked primary care records from one of the four catchment boroughs.

Outcome

The primary outcome of interest was the presence or absence of recorded physical pain within the clinical notes of the patients. This does not include mental, psychological or metaphorical pain.

Results

A total of 27,211 patients were retrieved based on the extraction criteria. Of these, 52% (14,202) had narrative text containing relevant mentions of physical pain. Patients who were older (OR 1.17, 95% CI 1.15-1.19), female (OR 1.42, 95% CI 1.35-1.49), of Asian (OR 1.30, 95% CI 1.16-1.45) or Black (OR 1.49, 95% CI 1.40-1.59) ethnicities, and living in deprived neighbourhoods (OR 1.64, 95% CI 1.55-1.73) showed higher odds of recorded pain. Patients with an SMI diagnosis were found to be less likely to report pain (OR 0.43, 95% CI 0.41-0.46, $p<0.001$). When comparing the overlap between primary and secondary care, 17% of the CRIS cohort also had records within LDN, and 31% of these had recorded pain in both records.

Conclusion

The findings of this study show the sociodemographic and diagnostic differences in recorded pain, and have significant implications for the assessment and management of physical pain in patients with mental health disorders.

Keywords: Natural Language Processing, Pain, Mental Health, Electronic Health Records

Strengths and Limitations of this study

- This study utilises natural language processing on clinical notes to access a large sample with information about pain.
- This is the first cross-sectional study to summarise and describe the distribution of recorded pain within the clinical notes of mental health records.
- The recorded mentions of pain within clinical notes clearly depend on the patient sharing and the clinician recording their experiences. When patients show no recorded pain, the study does not differentiate between pain that was discussed but not recorded, or pain that was not discussed.
- The findings are not generalisable to the general population since this study only looks at patients receiving mental healthcare within a specific geographic catchment.

9.3 Introduction

9.3.1 Background Rationale

Pain and its relationship with mental health are important research topics. Pain has imposed a significant burden on society in terms of medical care costs as well as lost productivity (Rayner et al., 2016). Pain is multifaceted, with physical, psychological, social, and biological causes and consequences (Merlin et al., 2014). Mental health disorders also present a considerable and complex public health problem, being a leading cause of disability and accounting for 28% of the national disease burden in the UK (Bridges, 2014). Electronic health records (EHRs) for mental health are a significant source of information for studying the

intersection between pain and mental health within those who receive specialist service input. EHRs open the possibility of investigating how pain is recorded and its impact on clinical outcomes.

Severe mental illnesses (SMIs) include diagnoses of schizophrenia spectrum disorder, bipolar disorder, or severe major depressive disorder (Abplanalp et al., 2020), where functional and occupational activities are severely impaired due to associated debilitating psychological problems (Public Health England, 2018). While several studies have looked at the relationship between pain and schizophrenia and bipolar disorders (Birgenheir et al., 2013; Bonnot et al., 2009; Potvin & Marchand, 2008; Stubbs et al., 2014) and at other mental illnesses such as depression (Bair et al., 2003; Blumer & Heilbronn, 1982; IsHak et al., 2018; Rayner et al., 2016; Thompson et al., 2016), the complex and potentially bidirectional nature of this relationship requires further understanding. Analysis of secondary data sources, such as EHR databases, might help by providing a fuller picture of the recorded clinical presentation of this group of patients; however, a prerequisite is that pain is adequately represented in derived data.

Demographic features such as age, gender and ethnicity can influence pain perception and experiences. Pain affects twice as many persons over the age of 60 as it does younger individuals (Noroozian et al., 2018)(Noroozian et al., 2018)(Noroozian et al., 2018). While pain is not a natural feature of the ageing process, many health conditions causing pain become more common with increasing age. Nonetheless, older patients often believe pain to be a normal aspect of ageing and might be hesitant when reporting it (Noroozian et al., 2018). There have also been variations found in the reported perception of pain by female and male patients, with female patients reporting experiencing more pain than males (Roger B. Fillingim, 2017; Vallerand & Polomano, 2000). Research has also shown disparities in pain

perception across different ethnicities, with individuals of Black (African) ethnicity reporting greater pain than White counterparts (Campbell & Edwards, 2012).

Socioeconomic status (SES) plays a role in health and overall well-being, with deprivation associated with unfavourable health outcomes and increased mortality rates (Martin et al., 2014). Patients with SMI already experience higher mortality rates than the general population, and this discrepancy is exacerbated by socioeconomic deprivation, primarily due to unequal access to good quality physical healthcare services (DE Hert et al., 2011; Frayne et al., 2005; Lambert & Newcomer, 2009; Laursen et al., 2009). Furthermore, patients with SMI continue to experience a decline in their SES over time, compounding its impact (Aro et al., 1995).

Most patient information is recorded in unstructured clinical narratives within EHR databases (Velupillai et al., 2018), and pain is likely to be no different, with few, if any, structured checklists ascertaining its presence in routine clinical care. Natural language processing (NLP), a computational approach to understanding and analysing human language, is therefore potentially useful for extracting such pain information. NLP has been applied extensively to EHR data, including studies of SMI, such as antipsychotic polypharmacy in mental health care (Kadra et al., 2018), multimorbidity in individuals with schizophrenia and bipolar disorders (Bendayan et al., 2022), and extracting symptoms of SMI (Jackson et al., 2017).

In addition to secondary care data, it is also useful to consider the recording of pain in primary care data. Within the UK, primary care is generally the first point of contact for patients (Sampson et al., 2015). Exploring the overlap of recorded pain between primary and secondary care could, therefore, provide a more comprehensive view of the patient's pain experiences, and any discrepancies could highlight gaps in care and communication.

9.3.2 Objectives

The objective of this study is to describe the distributions of recorded pain amongst mental health service users according to demographic factors such as age, gender and ethnicity, as well as neighbourhood deprivation levels and mental health diagnoses. This was achieved by examining recorded pain through the means of an NLP application within the clinical text of a mental health EHR database, and further evaluating this by measuring the overlap between pain recorded in secondary and primary health care, enabled through data linkage between the two.

9.4 Methods

9.4.1 Reporting

We use the RECORD (Benchimol et al., 2016) guidelines and checklist, an extension of the STROBE (von Elm et al., 2007) guidelines, for reporting the results of this study. This can be found in Appendix 7.

9.4.2 Setting

Data on recorded pain were obtained from the clinical text of a mental health EHR database, the Clinical Record Interactive Search (CRIS) resource. This contains a de-identified version of EHR data from The South London and Maudsley NHS Foundation Trust (SLaM), one of Europe's largest mental healthcare organisations (Stewart et al., 2009), which serves a geographic catchment of around 1.3 million residents in four south London boroughs (Croydon, Lambeth, Lewisham, Southwark). CRIS contains about 30 million free text documents, averaging 90 documents per patient (Velupillai et al., 2018).

Data were also obtained from a primary care database called Lambeth DataNet (LDN) (N H S, 2021a), which accesses all GP records from general practices based in the London borough of Lambeth. Data linkages (at the patient level) are already in place between CRIS and LDN (NIHR Maudsley Biomedical Research Center, n.d.).

Ethical Approval

CRIS and its associated linkages has received ethical approval as a data resource for secondary analysis from the Oxford C Research Ethics Committee (reference 23/SC/0257). A patient-led oversight committee (detailed in (Fernandes et al., 2013)) reviews and approves research projects that use the CRIS database. For service users, an opt-out system is in place and is advertised in all promotional materials and campaigns. Only authorised individuals can access this data from within a secure firewall. The CRIS project approval references for this work are 21-021 and 23-003.

LDN data access is overseen by the LDN Steering Group and the Caldicott Guardian acting for SE (south-east) London Clinical Commissioning Group. LDN approval was obtained as part of an existing CRIS project (project number 23-124) which included access to linked data from LDN (LDN project number 44, Caldicott Guardian approval, 15/9/21). This CRIS-LDN project aimed to examine the profile of patients with mental illnesses and chronic/persistent pain and compare them to controls from LDN who had chronic/persistent pain only.

Patient and Public Involvement

Patient and public involvement (PPI) in research is an active collaboration between researchers and members of the public, where the latter actively participate in contributing to the research (Research Design Service South Central, 2023). A PPI group with lived experiences of SMI and chronic pain were consulted as part of this research. The nature of

the data available was described to the group, and they were asked about their priorities regarding what research questions they would like answered. The group was unanimously interested in further study of the differences in pain experiences based on demographics and diagnoses. This was the main motivation for the objective of this study.

9.4.3 Participants

A cohort of patients was extracted from the CRIS database comprising those who were active (i.e., under an accepted referral) and aged 18+ on the index date of July 1, 2018, and whose record contained at least one document (>= 30 characters) within a window of July 1, 2017 to July 1, 2019. This window was chosen to avoid use of data collected during the coronavirus pandemic.

LDN extraction followed similar criteria for patients who were active on the index date, aged 18+, and contained pain diagnoses or medications from July 1, 2017 to July 1, 2019. Free-text information is unavailable within LDN, so no document criteria were required.

9.4.4 Variables

Demographics

Age, gender, and ethnicity variables were extracted from structured tables within the CRIS database. Individuals with missing gender or ethnicity values were retained as a separate category (Not stated/known). Ethnicity, in this context, encompasses both race and ethnicity but is referred to simply as ethnicity for the sake of simplicity.

Diagnosis

The primary diagnosis recorded closest to the index date of July 1, 2018, was extracted from the structured tables within the CRIS database. These are coded using ICD-10 (World Health Organization, 2008). The diagnosis codes were categorised as SMI (severe mental illnesses) and non-SMI, where SMI includes ICD-10 codes of F20-29 and F30-33.

Deprivation

The Index of Multiple Deprivation (IMD) decile measures from 2019 (GOV.UK, 2019) were extracted for information on neighbourhood deprivation for each patient, based on their address at the time of the index date aggregated by Lower Super Output Area (LSOA) - a standard national administrative unit containing an average 1500 residents. National Census data are used to calculate IMD scores for each LSOA. A lower IMD decile indicates higher deprivation levels. Individuals with missing IMD scores were retained in a separate category (Not known).

Recorded Pain

Pain-related keywords generated from a lexicon of pain terms (Chaturvedi et al., 2021), as described in Chapter 3, were used to identify patients in the cohort who had mentions of physical pain recorded in their clinical notes within the predetermined window. An NLP application was used on the documents of these patients. The application classified sentences within the document as relevant or not, where relevant refers to a mention of physical pain affecting the patient, and not relevant refers to no or negated mentions, hypothetical mentions, and metaphorical mentions of pain. These classes are the same as described in Chapters 5 and 6. Only relevant mentions were used in the results reported here. While the application classifies sentences, the outputs for each sentence classification was aggregated to patient level, where a single “relevant” output meant the patient had talked about relevant pain. If all the sentences for a patient were classified as “not relevant”, the

patient was considered to not have talked about relevant pain. The application has been described in detail in (Chaturvedi et al., 2023) and Chapter 6.

As with all other UK research based on access to anonymised primary care records, LDN does not allow access to any free text clinical notes. For this reason, pain information can only be extracted from the structured fields of the records. Read codes (NHS Digital, 2022b) were used to identify patients who had a pain diagnosis or were on any pain medications and treatments:

1. Pain medications code list - developed as part of a project described in (Ma, Romano, Ashworth, et al., 2022), which focused on analgesics (obtained from dm+d (a dictionary of medicines and devices (NHSBSA, 2023)) used in the treatment of 35 long-term conditions. These 35 conditions were obtained from (Barnett et al., 2012), a cross-sectional study on multimorbidities in patients registered with 314 medical practices in Scotland as of March 2007.

2. Pain diagnosis and treatments code list - developed as part of a collaboration project with Outcomes Based Healthcare (OBH), an organisation that provides a platform for the study of population health outcomes (Outcomes Based Healthcare, 2018), with the research described in (Hafezparast et al., 2023).

While these codes were developed for chronic pain, they are generic enough to be used for this research. These code lists are available on GitHub[24].

Anatomy Related to Recorded Pain

[24] https://github.com/jayachaturvedi/pain_in_mental_health

Another NLP application was developed for identifying anatomy mentioned in relation to pain. This was a classifier that generated a binary output - “mentioned” or “not mentioned”. This application was run on sentences labelled as relevant by the pain application. Once the sentences that contained mentions of body parts were identified, they were run through MedCATTrainer (Searle et al., 2019), which used named entity recognition (NER), a type of NLP task to label entities within the text to identify the specific body parts mentioned within the text. The purpose of using MedCATTrainer was that it linked the identified body parts to unique identification numbers (SCTID) from SNOMED CT, a terminology of clinical terms. These SCTIDs were used to aggregate the mentioned body parts, for ease of analysis. For example, foot, calf, and knee mentions would be aggregated under “lower limb”.

Overlap between CRIS and LDN

To examine the overlap across primary (LDN) and secondary (CRIS) care, the patient IDs from the CRIS cohort (N=27,211) were searched for matching records within the LDN database over the same window of July 1, 2017 to July 1, 2019. Variables were generated indicating the presence of the patients within LDN, along with variables indicating the presence of any codes for pain medication, diagnosis or treatment based on the predefined lists described above. This allowed the identification of patients with documented pain experiences in both their mental health and primary care records for the aligned time period. The cross-referencing process enabled the comparison of recorded pain between the two systems at the patient level.

9.4.5 Descriptive Statistics

All analysis was conducted using STATA v15.1 and the Python programming language (version 3.10.0).

Descriptive statistics were obtained for demographic, deprivation and diagnosis features and compared between the two groups - patients who had recorded pain (class 1, referred to as “relevant” in Chapter 5) and those who did not (class 0, referred to as “not relevant”, including negated mentions, in Chapter 5) - within their clinical notes. Logistic regression was conducted between the two classes to obtain unadjusted and adjusted odds ratios. Frequencies of body parts affected by pain and the overlap of recorded pain experiences between CRIS and LDN were also reported.

9.5 Results

9.5.1 Data Extraction

Based on the extraction criteria, 27,211 patients were extracted. Amongst these patients, 18,188 had pain keywords mentioned within their documents. These documents were run through the NLP application (SapBERT, as described in Chapter 6) to label them as relevant to pain (class 1) or not (class 0), resulting in 14,202 patients who had relevant mentions of pain within their clinical notes (Figure 9.1). While the NLP application classifies sentences as relevant or not, these sentence classifications are aggregated to patient level, where a minimum of 1 sentence classified as relevant will mean the patient has relevant mention of pain.

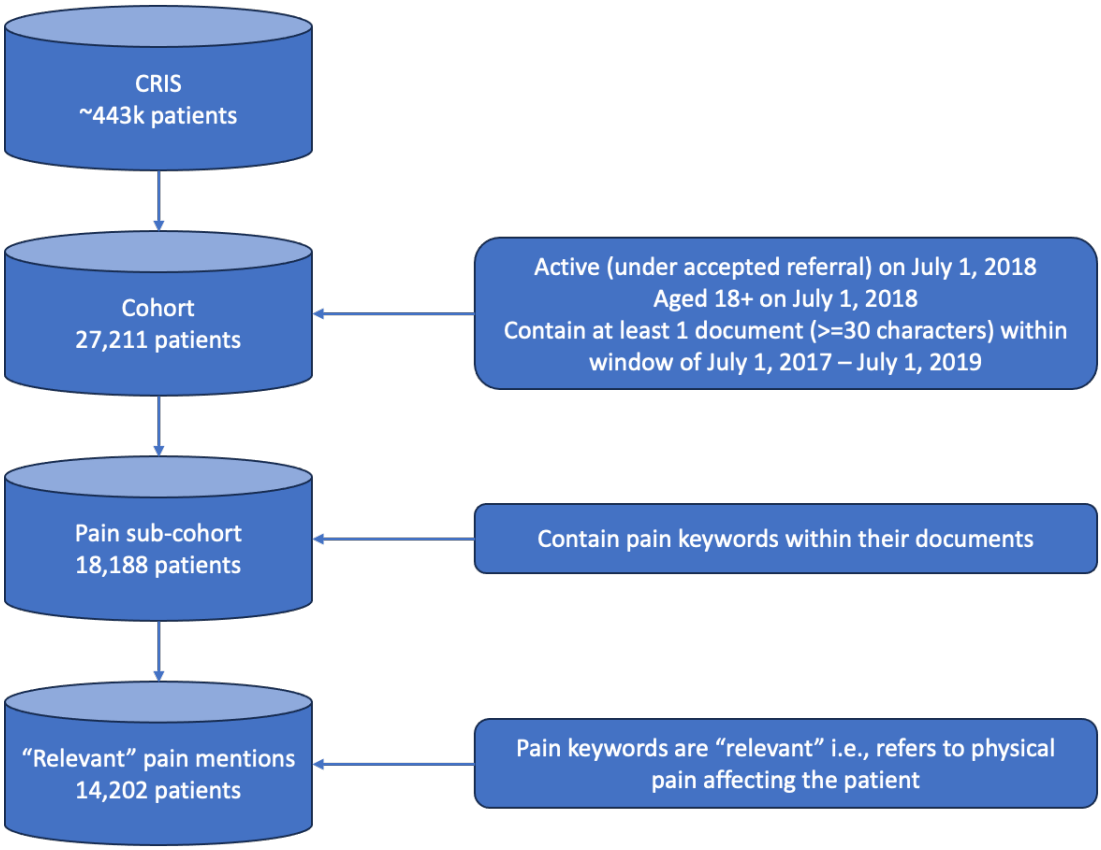


Figure 9.1. Data Extraction

This figure shows the process followed for the extraction of the cohort data, applying filters such as the patient being active i.e., under an accepted referral, aged 18+ and contain at least one document within a specified time period and index date.

9.5.2 Cohort Characteristics

Amongst the cohort of 27,211 patients, the mean age of the cohort was 44 (Inter-quartile range 29-55, SD 17.5), with 50.3% female and 48.2% of White ethnicity. The majority of the cohort (72.2%) lived in more deprived areas (IMD score <=5), and 67.0% received a non-SMI diagnosis. 66.8% of the patients (18,188 patients and 174,167 mentions within documents) contained pain keywords within their documents, and 52.1% of the cohort (14,202 patients) contained relevant mentions of pain in their documents.

9.5.3 Pain Mentions

Records of 52.1% of the patients within the cohort contained relevant mentions of pain. Differences between the patients who showed recorded pain (class 1) in their clinical notes and those who didn't (class 0) are shown in Table 9.1. Class 0 includes patients who did not have any pain mentions in their documents, as well as patients whose pain mentions were classified as not relevant. Patients within class 1 had an average of 10 pain mentions within their documents.

Characteristic	**n**	**Class 0 (no recorded pain)**	**Class 1 (recorded pain)**
N (%)	27,211	13,009 (47.9)	14,202 (52.1)
Mean Age (IQR)	44 (29–55)	41 (27–52)	46 (32–56)
Gender (N, %)			
Male	13,471	7,037 (54.1)	6,434 (45.3)
Female	13,709	5,953 (45.7)	7,756 (54.6)
Not known	31	19 (0.2)	12 (0.1)
Ethnicity (N, %)			
White	13,139	6,014 (46.2)	7,125 (50.1)
Black	5,866	2,115 (16.2)	3,751 (26.4)
Not stated/known	4,708	3,418 (26.2)	1,290 (9.0)
Asian	1,506	592 (4.5)	914 (6.4)
Other	1,197	512 (3.9)	685 (4.8)

Mixed	795	358 (2.7)	437 (3.0)
Index of Multiple Deprivation (N, %) Decile 2019			
<= 5 (more deprived)	19,660	8,847 (68.0)	10,813 (76.1)
> 5 (less deprived)	6,686	3,836 (29.4)	2,850 (20.0)
Not known	865	326 (2.5)	539 (3.9)
Primary Diagnosis: SMI vs Non-SMI (ICD-9 code) (N, %)			
SMI	8,962	3,059 (23.5)	5,903 (41.5)
Non-SMI	18,249	9,950 (76.5)	8,299 (58.5)

Table 9.1. Distributions between the two classes - class 0 (no recorded pain or not relevant) and class 1 (recorded pain or relevant)

This table summarises the distributions based on demographics, deprivation and primary diagnosis between the two classes (these classes are described in more detail in Chapters 5 and 6).

Demographic variations emerged between those with/without recorded pain in the cohort, as shown in Table 1. The mean age was higher in patients with recorded pain at 46 (SD=17) compared to 41 (SD=17) for the remainder. Patients with recorded pain were more likely to be female and had a higher representation across all ethnic minorities. Additionally, patients with documented pain experiences were more likely to live in higher deprivation neighbourhoods. In terms of diagnoses, SMIs were more prevalent in the recorded pain group, with ICD-10 chapter F20-29 (Schizophrenia, schizotypal and delusional disorders) being the most common diagnosis amongst class 1, accounting for 20.8% (95% CI 20.1-21.4) of the patients identified with mentions of pain, followed by 14.6% (95% CI 14.0-15.2) with the diagnosis of F30-39 (Mood [affective] disorders). Compared to the broader CRIS

population [31], females and ethnic minorities were over-represented among patients with pain mentions, while White patients were under-represented.

Table 9.2 presents demographic, deprivation and diagnostic associations with recorded pain obtained through logistic regressions (unadjusted and adjusted for different factors as detailed below).

	Logistic Regression Models				
	Unadjusted	Mutually adjusted			
		Model 1	Model 2	Model 3	Model 4
Age (per 10 years)	1.17 [1.15, 1.19] *	1.12 [1.11, 1.14] *	1.12 [1.11, 1.14] *	1.11 [1.10, 1.13] *	-
Gender					
Male	1 (reference)	1 (reference)	1 (reference)	1 (reference)	-
Female	1.42 [1.35, 1.49] *	1.42 [1.35, 1.49] *	1.43 [1.36, 1.50] *	1.47 [1.40, 1.55] *	-
Not known	0.69 [0.33, 1.42]	1.08 [0.50, 2.33]	1.06 [0.49, 2.30]	1.10 [0.51, 2.38]	-
Ethnicity					
White	1 (reference)	1 (reference)	1 (reference)	1 (reference)	1 (reference)
Asian	1.30 [1.16, 1.45] *	1.36 [1.22, 1.52] *	1.34 [1.19, 1.49] *	1.21 [1.08, 1.36] *	1.29 [1.15, 1.44] *
Black	1.49 [1.40, 1.59] *	1.58 [1.48, 1.69] *	1.50 [1.40, 1.60] *	1.25 [1.17, 1.34]	1.42 [1.33, 1.52] *
Other	1.12 [1.00, 1.27]	1.20 [1.06, 1.36]	1.17 [1.03, 1.32]	1.10 [0.97, 1.24]	1.08 [0.96, 1.33]
Mixed	1.03 [0.89, 1.18]	1.15 [0.99, 1.33]	1.12 [0.96, 1.30]	1.06 [0.91, 1.23]	1.01 [0.87, 1.17]
Not known	0.31 [0.29, 0.34] *	0.36 [0.34, 0.39] *	0.37 [0.34, 0.40] *	0.40 [0.37, 0.44] *	0.32 [0.30, 0.35] *
Index of Multiple Deprivation					
National Decile <=5	1.64 [1.55, 1.73] *	-	1.43 [1.35, 1.51] *	1.37 [1.29, 1.45] *	1.41 [1.33, 1.50] *

Diagnosis						
	SMI	0.43 [0.41, 0.46] *	-	-	0.56 [0.53, 0.59] *	-

Table 9.2. Logistic Regression findings for variables reflecting differences in class 0 (no recorded pain) and class 1 (recorded pain) (N = 27,211)

*Values are given as odds ratio (95% CI), and * indicates significance at p < 0.001*
Outcome is recorded pain vs no recorded pain.
Model 1 contained the demographic variables only [age, gender and ethnicity].
Model 2 contained the variables from Model 1, plus the variable for deprivation (IMD Decile].
Model 3 contained the variables from Model 2 plus the diagnosis variable.
Model 4 contains the ethnicity and deprivation variables alone.

Unadjusted odds ratios revealed patients with documented pain experiences were more likely to be older (OR 1.17, 95% CI 1.15-1.19, p<0.001), female (OR 1.42, 95% CI 1.35-1.49, p<0.001), of Asian (OR 1.30 in relation to a White reference group, 95% CI 1.16-1.45, p<0.001) or Black (OR 1.49, 95% CI 1.40-1.59, p<0.001) ethnicities, and living in deprived neighbourhoods (OR 1.64, 95% CI 1.55-1.73, p<0.001) when compared to the remainder of the sample. In a model containing all demographic variables (Model 1), the odds ratios were strengthened for all ethnic minority groups. Additional adjustment for neighbourhood deprivation (Model 2) resulted in a further strengthening of the odds ratio for females. In the model also adjusted for diagnoses (Model 3), odds ratios became stronger for females. Patients with SMI had lower odds of documented pain (OR 0.43, 95% CI 0.41-0.46, p<0.001) than non-SMI patients, with the odds ratio slightly weakening when adjusted for demographics, deprivation and diagnosis (Model 3). A supplementary model (Model 4) including both ethnicity and deprivation as covariates showed independent increased odds for Asian and Black patients and those in more deprived neighbourhoods.

9.5.4 Anatomy Distributions

Additional descriptive data were generated on the nature of the pain reported. Amongst the 14,202 patients with any recorded pain, there were 174,167 mentions of pain within the documents. Of these, 7,555 (53%) patients included 40,418 mentions of the anatomy associated with the pain. Of these 53%, each patient had an average of 5 body parts mentioned in the context of pain. The most common body part affected by pain, as per the recorded mentions, was lower limbs, which accounted for 20% of all mentions where anatomy could be ascertained (Table 9.3).

Body Part	Mentions	Frequency (mention-level)
Lower limbs	Feet, ankle, leg, knee, calf, thigh, toes	20%
Upper body, excluding back	Chest, side of chest, upper body, torso	19%
Upper limbs	Hand, wrist, arm, elbow, thumb, shoulder	17%
Stomach/abdomen region	Stomach, abdomen, groin, bladder, prostate	16%
Head and neck	Head, tooth, face, mouth, tongue, eye, ear, neck	15%
Non-specific site	Entire body, skin, muscle, joint	8%
Back	Back, lower back	5%

Table 9.3. Body parts affected (at mention level)

Body parts have been aggregated for ease of summarisation.

9.5.5 Overlap with Primary Care

When comparing secondary care CRIS records with those of primary care from LDN, among the 27,211 patients of the CRIS cohort, 4,822 patients (17%) also had records in LDN. Amongst these patients who had records in both CRIS and LDN, 1,507 (31%) patients were identified as having some recorded instance of pain in both their records, while 687 (14%) patients showed recorded pain only in LDN (primary care). Among the 27,211 patients within CRIS, 12,695 (46%) had recorded pain only within CRIS (mental health care), as seen in Figure 9.2.

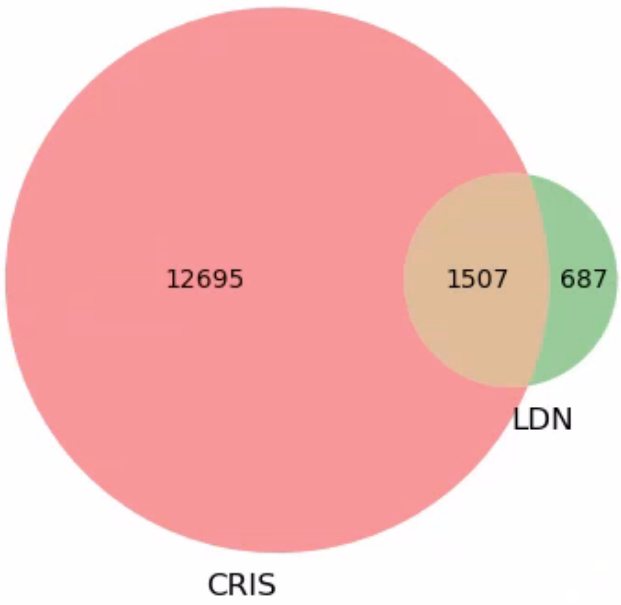


Figure 9.2. Overlap of recorded pain between CRIS and LDN
This Venn diagram shows the overlap of patients with recorded pain within CRIS and LDN. The Venn diagram is unweighted to keep in proportion with the amount of overlap, for visual clarity.

9.6 Discussion

This study investigated the differences observed in recorded pain mentions within the clinical notes of mental health records. The results reflect current literature findings that pain is a common issue among patients with mental health disorders. In a cohort of 27,211 patients, 18,188 (67%) patients contained pain-related keywords in their text, and 14,202 (52%) patients had relevant pain mentions, i.e., the mention indicated physical pain affecting the patient in question, as determined by the NLP application. Disparities in recorded pain mentions were found across genders, with females being over-represented. This is consistent with other research that indicates gender disparities in pain experiences (R B Fillingim, 2000; Vallerand & Polomano, 2000; Wandner et al., 2012). Furthermore, while patients with known ethnicities were mostly over-represented in the cohort of relevant pain mentions (in relation to those with unknown ethnicity), most noticeable were the Black, Asian and other ethnic groups. This aligns with research around the undertreatment of patients within certain ethnic minority groups (Green et al., 2003) and highlights the need for a comprehensive exploration of pain experiences across diverse populations. Moreover, the study's findings are also consistent with studies that indicate the impact of deprivation on health outcomes (Martin et al., 2014), as people living in more deprived areas (IMD decile <= 5) were more frequently recorded with pain.

When comparing the overlap of patients between primary and secondary care, it was found that 17% of the patients within the CRIS cohort also had records within LDN. Amongst these patients, 31% had recorded pain instances in both records. While this overlap between primary and secondary care seems low, it is important to bear in mind that Lambeth only represents 22% of the catchment covered by CRIS (Perera et al., 2016). Patients present in CRIS but not in LDN could include patients who have recorded instances of pain within the

free-text clinical notes in LDN and might have been missed in this study since we do not have access to this text. Furthermore, this study did not differentiate between acute and chronic pain mentions and focused on extracting mentions of physical pain of any duration. As a result, the higher occurrence of pain mentioned within CRIS can be partially attributed to the documentation of such acute or short-lived pain episodes. Conversely, the GP records within LDN likely focus on recording persistent and chronic pain experiences. This disparity in recording pain should be considered when interpreting the findings of this study. Looking specifically at chronic pain instances within the CRIS notes may improve the comparability. However, the temporal information required to determine pain chronicity from clinical notes is a particular challenge and can be difficult to extract reliably. Future work can attempt to differentiate acute and chronic pain through temporal or contextual information, which could provide richer insight. However, the current broad inclusion of pain provides wider coverage for this initial exploration of pain mentioned within clinical notes.

A strength of this study is the size of the data set available and the access to information about pain from the clinical text. To the best of our knowledge, this is potentially the first cross-sectional study to summarise and describe the distribution of recorded pain derived from routine mental health records. While the cohort data extraction did not apply any filters on demographics, aiming for broad representativeness, other systemic biases related to access to healthcare resources may still exist. Factors like deprivation level and ethnicity can influence the utilisation of services and, therefore, documentation within health records, often stemming from perceived barriers to access. However, by not restricting cohort selection on demographic factors, this study intended to capture a diverse patient population receiving care across the South London boroughs. In addition, another strength of this project lies in the development and application of the novel NLP approach, which facilitated the extraction of this pain information from unstructured clinical notes. Unlike traditional methods that rely

solely on structured data fields or manual chart reviews, the NLP application designed for this study enabled the automated extraction of this information from extensive free-text clinical narratives at scale. This innovative use of NLP allowed for the identification and extraction of pain-related mentions that may have been otherwise overlooked or challenging to capture through conventional means. By leveraging the rich contextual information embedded within the clinical notes, the NLP application could accurately classify instances of pain. Moreover, the integration of domain-specific knowledge into the SapBERT model, which was used here, further enhanced the model's performance.

A limitation of this study is that the recorded mentions of pain within clinical notes depend on the clinician recording them. The actual occurrences of pain experiences could remain unaccounted for if they weren't recorded by the clinicians or were not shared with the clinicians, especially for patients with severe mental illnesses who might be completely or partially nonverbal. While the NLP application achieved good performance metrics during its development and evaluation, it is not impervious to imperfections. Instances of pain experiences might have been overlooked if they were not included as examples during the training of the application.

The scope of this study is limited to the examination of mental health records from an EHR database in South London. Given the absence of a comparative cohort of patients experiencing pain without any mental health disorders, the findings of this study are not generalisable to the overall population. However, they might be relevant and generalisable to some extent to other populations of patients with mental health disorders. It is essential to acknowledge the potential influence of gender and ethnicity on the reporting of pain experiences, particularly if females and minority ethnicities (due to language barriers or other reasons) are less likely to self-report their pain experiences (Green et al., 2003; Hoffmann &

Tarzian, 2001; Samulowitz et al., 2018). Since the focus of this study has been on a mental health EHR database, the clinical care within this setting is focused on mental health issues reported by the patients. Consequently, as much importance might not be given to the investigation and reporting of physical health conditions such as pain.

This study cannot determine a cause-and-effect relationship or directionality between pain and mental illnesses. Despite this, the study has highlighted existing disparities in recorded pain experiences and brings to attention the need for further research to better understand and address them at the point of care.

9.7 Conclusion

The outcomes of this study have significant implications for the assessment and management of pain amongst patients with mental health disorders and highlight the importance of utilising NLP methods on EHR databases for research purposes. Notably, these findings reiterate the recommendations set forth by Mental Health America (American Psychiatric Association, 2020), advocating the need for proactive initiation of conversations around mental health and pain with patients. Relying solely on patients to self-report symptoms could potentially lead to worse outcomes, especially since the stigma surrounding pain and mental health conditions may prevent patients from seeking the necessary treatment. Thus, early and proactive interventions could go a long way towards improved long-term outcomes. Unfortunately, there still exists a perceived lack of credibility and empathy towards patients living with pain (Bennett, 2020), particularly when compounded by co-existent mental illnesses. This was one of the main points shared by the PPI group consulted as part of this study. More research in this area can help towards these issues and provide safer and equitable access to good-quality pain management.

While these findings represent a step forward, they are only one side of the story. Combining these findings with patient-reported insights could offer a more comprehensive understanding of pain experiences within this cohort. However, achieving this is a challenging task due to the lack of such data and the inability to link patient-reported experiences to their health records. Further research is needed to better understand the relationship between pain and mental health and to develop more effective interventions to manage pain in this population.

9.8 Data Availability Statement

Data are owned by a third party, Maudsley Biomedical Research Centre (BRC) Clinical Records Interactive Search (CRIS) tool, which provides access to anonymised data derived from SLaM electronic medical records. These data, and the NLP application, can be accessed by permitted individuals from within a secure firewall (i.e. the data cannot be sent elsewhere) in the same manner as the authors. For more information, please contact cris.administrator@slam.nhs.uk. Any STATA and Python code used in this project will be available on GitHub[25].

9.9 Author Contributions

The idea was conceived by JC, AR, and RS. JC conducted the analyses and drafted the manuscript. MA provided insights on LDN data. AR and RS provided guidance in the design and interpretation of results. All authors commented on drafts of the manuscript and approved the final version.

[25] https://github.com/jayachaturvedi/pain_in_mental_health

9.10 Funding and Acknowledgements

AR is funded by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. RS is part-funded by i) the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; ii) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; iii) the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386); iv) the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities. JC is supported by the KCL-funded Centre for Doctoral Training (CDT) in Data-Driven Health. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. RS declared research support received in the last 36 months from Janssen, GSK and Takeda. All other authors declare no other competing interests.

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 and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the 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.

This work uses data provided by patients and collected by the NHS as part of their care and support. An application for access to the Clinical Record Interactive Search (CRIS) database for this project was submitted and approved by the CRIS Oversight Committee. The authors would like to acknowledge Dr Ruimin Ma for her help in obtaining the LDN codes.

CHAPTER 10: Conclusions and Future Work

Having reviewed the background literature and stated the motivation of this thesis in Chapters 1 and 2, the aims of this thesis were to develop a clinical NLP application to classify sentences as containing mentions of pain or not, linking the pain entities within the clinical text to SNOMED CT modelled as a knowledge graph, and conducting experiments to test the impact of such incorporated knowledge on the predetermined classification task. To address these aims, four sentence-level classifiers were developed - two baseline models that did not incorporate external domain knowledge (Random Forest and BERT_base), described in Chapter 6, and two models that incorporated external domain knowledge (SapBERT and Random Forest with KGE), described in Chapters 6 and 7. Their performance was compared, and error analyses were conducted (Chapter 8). These classifiers were then applied to sentences from a patient cohort to compare the differences in output (Chapter 8). The classified sentences from the best performing model, SapBERT, were used in a prevalence study to understand the distributions of recorded pain experiences based on demographics, diagnoses and deprivation (Chapter 9). The findings from this thesis provide insights into the advantages and limitations of NLP models that incorporate external domain knowledge for the identification of relevant pain mentions within unstructured clinical notes. The aims and the research questions answered through this project are discussed in more detail below. Most of the preceding chapters include their own discussion and conclusion sections. This chapter supplements them and adds information on how the aims were achieved and research questions answered, and describes potential future work.

10.1 Aims achieved

Aim 1: Development of an NLP application to classify sentences as containing mentions of physical pain or not.

This aim was achieved through a multi-step process. First, a lexicon of pain terms was developed to identify documents within the CRIS database that mentioned any pain-related terms. Chapter 3 details the iterative development and validation of this comprehensive lexicon. This lexicon was used in the extraction of data for the development of gold standard annotations that were used to train the classifier models. The data extraction process is described in Section 4.12.3 of Chapter 4. Next, a subset of extracted notes was manually annotated by medical students to create a gold standard for model training and evaluation, as discussed in Chapter 5. Four classifiers were developed (Random Forest, SVM, KNN, BERT_base) and trained on the gold standard labelled data, with hyperparameters tuned for optimal classification performance. Chapter 6 provides in-depth comparisons of model architectures, parameters, and evaluation metrics. Amongst these four models, the BERT_base model achieved the highest performance, with its transformer-based neural architecture enabling it to learn contextual representations better than the other models. This aligns with the previously discussed importance of context for a better understanding of pain mentioned within the clinical text, and this model performing best is evidence to this point.

Aim 2: Linking pain entities within the clinical text to compositional structured knowledge, such as SNOMED CT modelled as knowledge graphs, to make use of the additional relations between the concepts for better utilisation of such information in the classification of sentences within the clinical text.

This aim focused on linking pain entities extracted from clinical text to structured knowledge modelled as knowledge graphs with the objective of enriching the model with such external domain knowledge and ultimately improving classification performance. This aim was achieved through two approaches. Firstly, the SapBERT model, which integrates UMLS during its pretraining, was fine-tuned on the gold standard annotation data. Chapter 6 details SapBERT 's architecture and compares its performance to other classifiers. In addition to this, a KGE model was constructed by combining the pain mentions within the gold standard annotations, treated as subjects, with their corresponding SNOMED CT concepts as subject-predicate-object triples, as described in Section 7.4 of Chapter 7. Furthermore, this KGE model was then incorporated into a Random Forest classifier, with the parameters and performance metrics described in Section 7.9 of Chapter 7**.** Both the models that incorporate knowledge, SapBERT and Random Forest with KGE, achieved similar performance metrics and had similar frequencies of class proportions upon being run on sentences from the cohort for the prevalence study. In addition, incorporating domain knowledge into the Random Forest model made this standard ML algorithm almost as good as the transformer-based models. The results demonstrate the feasibility of augmenting clinical NLP with external domain knowledge resources.

Aim 3: Conduct experiments to test the impact of such incorporated knowledge, compared to baseline NLP methods, using NLP-derived information in a downstream epidemiological study.

This aim involved comparative experiments to assess the real-world impact of incorporating external knowledge into clinical NLP methods. The four classifier models - two baselines and two incorporating knowledge - were applied to sentences from a patient cohort extracted from the CRIS database. Overlap between them was compared, and there was lower inter-model

agreement, indicating differences in classification despite similar performance metrics and class proportions. These comparisons are detailed in Chapter 8, and highlight the potential for ensemble methods where the strengths of the different classifiers can be exploited. Additionally, a downstream prevalence study was conducted using the predictions from the best performing model, SapBERT, on the cohort to better understand the distribution of recorded pain. This epidemiological analysis provided insights into recorded pain distributions across demographic and clinical characteristics, as described in Chapter 9. As part of external validation, as well as to study the continuity of care, the cohort from CRIS was also compared to an external dataset from LDN to understand the overlap in recorded pain between primary and secondary care, which is also described in Chapter 9.

10.2 Research Questions Answered

These aims tie into the research questions posed in Section 1.1.4 of Chapter 1 at the beginning of this thesis.

Q1. Does a system that incorporates domain knowledge into an NLP task perform better than a system without that knowledge?

By comparing the classifiers that incorporated external knowledge to those that did not, it was found that the former did perform better. The F1-scores were higher for the models incorporating knowledge. Moreover, the error analysis revealed that the models incorporating knowledge were more successful at dealing with common misclassifications. This result is consistent with previous research where external domain knowledge was incorporated into classification tasks and outperformed various existing models on general domain (non-medical) benchmark datasets (Ennajari et al., 2022).

Q2. Does this method successfully harness the relations that exist between the pain concepts within a structured knowledge resource and translate them to better classification of sentences within the EHR text?

When comparing the method in which the different models tokenised the sentences, it was noticed that SapBERT, which is pretrained on UMLS, performed tokenisation more efficiently due to access to the medical vocabulary from UMLS. I hypothesised that this could have led to better entity detection and, thereby, better classification performance. In addition to this, the random model forest that used embeddings from the KGE model also outperformed its counterpart random forest model without any external knowledge. This indicates that the additional information about the triples obtained from the SNOMED CT KG and the KGE model led to better identification of features for the classes and, therefore, better classification performance. SNOMED CT KGEs have previously demonstrated proficiency in tasks like relation prediction and entity classification (Chang et al. 2020). The findings presented here further extend their success, showcasing their effectiveness in sentence classification as well.

Q3. Can this approach be harnessed to extract richer information about pain from mental health EHRs, thereby improving the quality of research outputs?

The best-performing model, SapBERT, which did incorporate external knowledge, performed the best amongst all the classifier models and was therefore used to identify mentions of pain within sentences of a cohort extracted from CRIS. In addition to this, the anatomy classifier, also built by fine-tuning SapBERT, described in Section 6.9 of Chapter 4, could accurately identify pain sentences that also mentioned anatomy associated with pain.

10.3 Strengths and Limitations

A major strength of this work is the novel approach of incorporating external structured knowledge into classical NLP approaches, thereby combining the logic-driven approach of structured knowledge with that of the statistical, data-driven approach of classic NLP, and the application of this method to a use case of pain in mental health records. Additionally, the annotated corpus created through the exhaustive annotation process represents one of the largest collections of mental health documents that include mentions of pain. This high-quality labelled dataset can facilitate future research and benchmarking. In addition, I believe that the lexicon developed for pain, leveraging multiple resources, is the most comprehensive one available and publicly shared. Finally, the secondary analysis integrating primary and secondary care records provides a more complete picture of pain documentation across care settings. This showcases the value of leveraging data from multiple sources.

A critical component that enriched this work was the engagement of a PPI group consisting of individuals with lived experience of chronic pain and severe mental illnesses. The first-hand perspectives of this group helped guide important decisions like the analysis conducted within the prevalence study. This ensured the work aligned with factors most meaningful and impactful to those directly affected. By collaborating directly with people suffering from these conditions, the research questions, process, and findings were refined to be more practical and relevant. The PPI involvement exemplified the benefits of true partnerships with patients to produce more meaningful work. This collaborative approach should serve as a model for future efforts to tap into patient experiences and values, especially when researching something as subjective as pain.

However, a limitation when applying NLP to any clinical text is that documentation is dependent on what the clinicians record. The absence of a mention cannot definitively indicate the absence of a condition or symptom in a patient. It simply means the details were not captured in the written notes, or were not shared by the patient, which could be due to various factors. So, the results here reflect what was documented rather than actual prevalence. Another limitation is the substantial computational resources required for knowledge-based NLP methods compared to baseline approaches. The knowledge graph integration and enhanced contextual modelling increase the model size and training times. This could hinder larger-scale implementation. Nevertheless, the utilisation of batch processing to improve run time and optimization of compute methods for better memory usage can mitigate these issues.

In retrospect, a few improvements could have strengthened the methodology. The annotation process required iterative refinement of guidelines due to ambiguity around capturing the nuances of the pain mentions within the clinical notes. Rather than forcing binary decisions on complex cases that caused disagreements amongst the annotators, retaining disagreed instances and training the model to predict a probability distribution over classes may have better represented realistic documentation patterns. Similarly, negated mentions could have been used as a separate category given their distinction from the not relevant class of sentences. Follow-up efforts could implement a multi-class architecture explicitly delineating between the negated and not relevant classes. Regarding integration of structured domain knowledge into the classifier models, applying recent state-of-the-art attention mechanisms could lead to better performance and prove to be a progressive iteration upon the strategies used here. While the approaches undertaken in this project provide a good foundation, exploring new techniques and architectural decisions could take this task to the next level.

In summary, this work pioneered an NLP methodology that leveraged external structured knowledge and generated useful resources, providing insights into recorded pain experiences in an understudied population. Nonetheless, continued research is needed to address reliance on clinical documentation practices and optimization of advanced NLP techniques.

10.4 Contributions and Impact

10.4.1 Accessible to the Community

In the spirit of keeping my research openly available to the community, I developed a website[26] that included details and showcased findings from the prevalence study, as well as any other material that was presented at conferences and events. The link to this website has been made available as a QR code on all my presentations and posters. Since its creation in May 2022, 245 people have visited the website, with the peaks in visits and locations coinciding with conferences and other social media promotions (Figure 10.1).

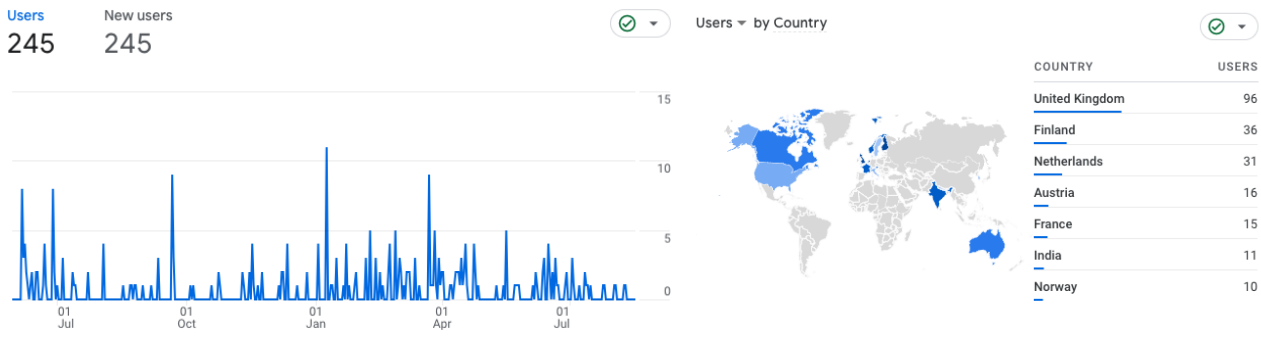


Figure 10.1. Google Trends for the website since its creation
This figure shows the impact of the project website by looking at usage over time since its creation, showing frequency of use on the left hand side and distribution of users by country on the right hand side.

[26] https://sites.google.com/view/pain-mental-health/

The website and research were promoted by Mental Health Research Incubator[27], one of a series of incubators established by the NIHR to build research capacity in priority areas. I also had the opportunity to write a blog[28] for the NIHR, which was widely promoted on social media.

10.4.2 Incorporation into MSc Dissertations

I have collaborated with two MSc students and co-supervised them on their dissertation topics which were focused on the theme of pain. One of them leveraged social media data by looking at subreddits about pain and conducted analysis on the different topics discussed within these forums as well as changes in sentiment over time. The most common topics discussed were physical symptoms, emotions, and peer advice. No difference was found in sentiments over time. The other student used MIMIC-III to identify patients experiencing pain within the database, by developing classifiers that incorporated features such as type and quantity of analgesics prescribed, readmissions, and mentions within the text. Chest pain was the most frequent pain mentioned, and aspirin was the most common analgesic prescribed. The best-performing classifier achieved an F1 score of 0.95. Both students used the pain lexicon to help them identify relevant posts and documents about pain.

10.5 Future work

This research holds significant potential for a range of future applications, offering the opportunity to enhance our existing knowledge of pain. By making all the code openly accessible, the aim is to facilitate broader development and adaptation for implementation across various mental health EHR databases in different Trusts, as well as other types of

[27] https://mentalhealthresearch.org.uk/studies/pain-and-mental-health/

[28] https://www.maudsleybrc.nihr.ac.uk/posts/2023/august/identifying-mentions-of-pain-in-mental-health-records-text-a-natural-language-processing-approach/

EHR systems. A cross-evaluation with additional mental health trusts will provide valuable insights into the models' reproducibility. Assessing the performance of the applications on the openly available MMIC-III dataset (intensive care unit database) will be of particular interest to better understand the reproducibility on different types of data. Beyond its primary purpose in epidemiological research for identifying patient cohorts experiencing pain within a database, there are clinical applications as well. This tool can support clinical decision-making by offering insights to clinicians about groups that may be vulnerable to pain based on various sociodemographic and diagnostic factors. Additionally, it holds potential in public health monitoring, allowing for a more comprehensive understanding of the population's needs using EHR data. This would be an improvement upon the current situation, where the available information in structured fields is somewhat restricted or limited in scope. Finally, while the models were developed for the use case of pain, this process can be replicated for other clinical entities.

Details of some upcoming work that will directly build upon what has been developed so far are provided below.

10.5.1 Pain Ontology

The lexicon of pain terms will be formalised into an ontology for ease of use by other researchers in the community. BioPortal will be the platform for hosting this ontology. Some preliminary work by Smith et al. (2011) provides a general structure for an ontology of pain and suggests a categorisation of pain that might be beneficial to follow in the ontology (Smith et al., 2011), as shown in Figure 10.2.

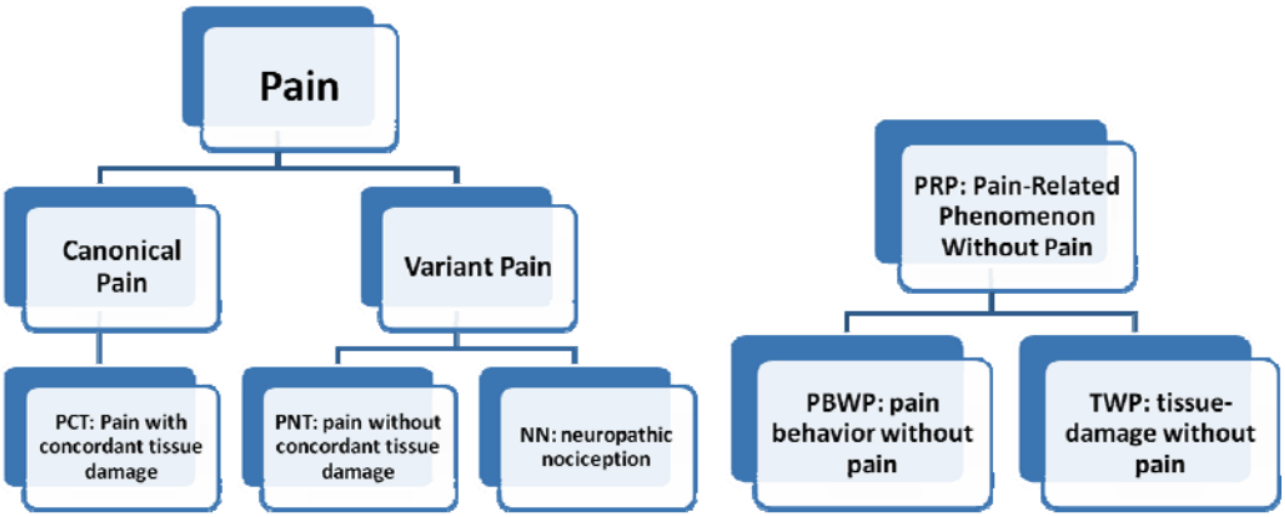


Figure 10.2 Categorisation of pain by (Smith et al., 2011)
This flow diagram shows a recommendation of the categorisations of pain by Smith et al. (2011) which could be adapted into the construction of the pain ontology based on the pain lexicon described in Chapter 3.

10.5.2 CRIS Deployment

The pain classification application, incorporating SapBERT, was prepared for deployment on the entire CRIS database, where it will undergo further manual validation and then be made available to other researchers who utilise the database. The NLP application was submitted to the CRIS team by providing them with the model and an NLP application document[29] that details what the application does, how it performs on test sets, and its expected outputs.

The process followed for deployment over all of CRIS is as follows:

The trained machine learning model is packaged into a Docker[30] container, which is a standardised software unit containing all necessary dependencies and configurations to run

[29] https://docs.google.com/document/d/13bdNCcQO6H3Gqrd5Z-eg-sYc-Hvii1TH/edit?usp=sharing&ouid=104056526702938298903&rtpof=true&sd=true
[30] https://www.docker.com/

the model. This packaging into a container eases deployment across different environments. The Docker container is then pushed to a repository of containers that interfaces with the cloud machine and local databases. Metadata like the container location, database column names, and model outputs (predicted class, probability, keywords) is uploaded as a .zip file onto the NHSTA (NHS Text Analytics) cloud where all the applications are stored. A pipeline on the NHSTA platform takes new clinical data from the CRIS system, sends it to the Docker container for classification by the NLP model, and writes the results back into the CRIS database tables. Currently, the predictions from the deployed model have undergone one round of manual validation on 250 sentences, achieving a precision of 88% and a recall of 78%. The application has been deemed ready for deployment and will be run on the entire CRIS database on a monthly and ad hoc basis, making the outputs of this application available to other researchers.

10.5.3 Unanswered PPI Questions

The discussions with the PPI group led to a long list of research questions that could be answered by access to such pain data. The main PPI group question was around the differences in the distribution of pain experiences based on demographics and diagnosis and differences in body parts affected by pain. While this has been addressed in this thesis, other questions remain unanswered. Despite the fact that not all of these questions can be answered by the data available, they are important and will be under consideration in future research projects. These outputs from the PPI meeting will also be made available in the public domain, in the form of a CRIS blog, to highlight these pain-related research priorities. The questions are:

1. Somatisation of pain in patients with SMI and depression

Somatisation refers to the manifestation of psychological distress by the presentation of physical symptoms. An approach to answering this question could be to look for keywords that could indicate somatisation within the clinical notes. While somatic pain is a term included in the pain lexicon, there might be a need for a more specific search within these documents to identify more examples of how somatisation is discussed within the notes. Another classifier could be built using these specific examples to classify sentences as mentions of somatic pain or not.

2. Effect of not getting a pain diagnosis

   Multiple members of the PPI group had concerns about this topic, as they personally experienced delays in getting a proper diagnosis that explained their pain. They were curious as to whether a patient's pain gets better once they receive an official diagnosis for their pain. Their justification for this concern was that when one doesn't know what the cause of their pain is, the pain tends to feel more amplified. In addition to this, they wondered if patients' mental health could get worse when they don't receive a diagnosis for their pain. An approach to answer this could be to identify when a patient receives a pain diagnosis and compare mentions of pain before and after this diagnosis date. The hypothesis would be that pain mentions would reduce after receiving a diagnosis.

3. Emotional pain vs. physical pain

   The interest in this stemmed from a need to understand and demonstrate the mental element and its effect on pain. While this current work did not include instances of emotional and mental pain, a separate classification could be undertaken to identify these mentions and use the results to study the effects and distributions of emotional and mental pain.

4. Patient words vs. Clinical words

   A limitation of this work, and any work that relies on clinical notes, is that the information recorded is from the clinician's perspective and their interpretations of what the patient is experiencing. With pain being such a personal and individual experience, not considering the patient's perspective can be detrimental. An approach to incorporate this element might be to find quoted text within the notes, which might encapsulate the patient's words. However, an initial exploration of the data might be required to understand how frequently clinicians document patient words in quotations. An application exists for the identification of quoted text within CRIS, which can be harnessed to help answer this question.

5. Effect of recreational drugs vs. medical drugs on pain

   The group were interested in knowing the impact of using recreational drugs such as marijuana, as well as measures like meditation and focus on pain memory, on pain experiences. They also expressed interest in further research into the effects of SMI medications on pain. Some SMI medications are believed to lead to side effects such as weight gain which can worsen painful conditions like joint pain and can lead to worse outcomes. The CRIS database has access to NLP applications for the identification of recreational drugs within the documents, as well as the identification of various classes of medications using their medication application. These NLP applications can be utilised to answer these questions.

6. Believability

   A major concern amongst the PPI group was the believability of their pain, given their mental health diagnosis, and the effect that has on the clinicians' judgement and

preconceptions about their pain. They mentioned the term “frequent flyer” that they believe is used against them due to their need to sometimes have multiple hospital admissions due to pain, or other reasons. It would be interesting to explore whether something like this would be recorded within clinical notes, and if clinicians refer to certain patients as “frequent flyers” or “revolving door” with the notes, and if mentions of such words can suggest the preconceptions of the clinicians.

7. Coping

   A final question the group was interested in was about how patients with different mental health diagnoses were living and coping with pain. It would be challenging to identify this within clinical notes, as coping can vary from patient to patient. However, words like “struggle”, “cope”, “can’t do”, etc. might indicate something.

This thesis only uncovers a small portion of research around pain and mental health and has the scope to expand in various directions and provide more information to other researchers and clinicians and have an impact on the care provided to patients living with pain.

11 Appendices

Appendix 1 - Decision Logbook

Decision	**Justification**
50 randomly selected documents were used for pain exploration in CRIS, mimic, Twitter and Reddit	The number of documents was limited to 50 for pragmatic reasons: manual review is a labour-intensive process, and the resources were not available to review more than 50 per source. This decision did not impact the lexicon development, as these documents are used only for exploration, with embeddings built on the whole of two sources (MIMIC and CRIS) used to generate the terms for the lexicon) to supplement the development of the lexicon.
3 annotators	Previous literature (Sim & Wright, 2005) stated that in order to achieve a Cohen's kappa of greater than 0.40, it is not advantageous to use more than 3 raters/annotators, as it does not make much more difference to the agreement measures. While two annotators might have been sufficient, three were available and helped develop more gold-standard annotations than might have been possible with two.
When selecting the span of the sentence	The length of sentences was calculated to get the average length and use that to determine how many characters would

for training the classifier and running the predictions by the classifier, 200 characters were selected before and after the pain keyword.	be sufficient to cover a sentence, which will thereby be enough context for the keyword. The average length of sentences was 200 characters.
No filters were applied when extracting documents for annotation, such as diagnosis or age.	This was so that the developed application would not be biased to a particular diagnosis or age group and be built on a random sample from the CRIS database.
Why was UmlsBERT not used when it was similar to SapBERT in its development?	UmlsBERT was the first choice, but a lot of their available code was deprecated and not actively updated by the authors. Upon contacting the authors (by email and posting issues on GitHub), no response from them forced me to move on to an alternative.
Why was machine learning used in this project rather than pattern matching?	There was no discernible pattern with the mentions of pain for me to generate rules-based pattern matching.

Why was AmpliGraph used?	It was easy to implement and actively maintained, with a Slack channel for support. Other alternatives were more complex, and the creators were not easily reachable to help with technical issues.
Why was sentence-level classification done instead of document level?	Sentence level allows for focus on precision and capturing every instance of pain. It can easily be aggregated to get document-level classes, while vice versa is not possible.
GP records - LDN	They are another dataset similar to CRIS with a known overlap of patients, and so they were a good source for external validation.
10-fold cross-validation	A value of k=10 is very common and highly recommended in the field of applied machine learning. A value of k=10 has been shown empirically to yield test error rate estimates that neither have excessively high bias nor very high variance (James et al., 2013).
Confidence intervals in performance metrics	A lot of NLP literature compares various NLP models and highlights improvements in performance by small margins. Being part of the Biostatistics department, there was some disagreement on this NLP practice of not including confidence intervals with our performance metrics. In order to

	address these concerns and be more robust in the reporting of such metrics, I have included confidence intervals for all the metrics reported in this thesis. They were calculated by utilising bootstrapping approaches.
Index date and window for the prevalence study	Upon discussion with the CRIS team, it was identified that using a date that was halfway through the year was best as it avoided any seasonal differences that have an impact on attendance. Also, since COVID-19, there might have been variations in the data, such as remote consulting, and so it was best to avoid any dates from 2020 onwards. For this reason, July 1, 2018 was chosen as the index date, and a period not affected by COVID was used as the window (2017-2019).
Why did I not utilise data from Kings College Hospital (KCH)?	Access to data from KCH requires having an existing principal investigator (PI) based at the hospital. I got in touch with some potential PIs and had some preliminary talks with researchers from the Cicely Saunders palliative care department, but since my project is more generic i.e., not focused on a particular department, and it was too late in my project to change this, any collaborations with them did not work out. In the end, I was unable to find a PI at KCH.

Why did I not focus on chronic pain?	From preliminary explorations, it was evident that it would be hard to determine temporality in order to determine chronicity with enough confidence. A first round of annotations was conducted where temporality was included as one of the annotations, but there were too many disagreements in how the annotators interpreted the temporality, so it was removed.
Why did I not look at the Brief Pain Inventory form at SLaM?	This form is not included within CRIS but is a part of the ePJS (Electronic Patient Journey System), and so an additional linkage would have to be developed. There were some time delays in figuring out whether developing such a linkage was feasible as the person responsible did not have access to the back end of the database, and the person eventually left. This was during the end of 2020-early 2021, and it was hard to coordinate efforts towards this remotely. In the end, it did not pan out.
Why did I not look at IAPT chronic pain clinic data?	IAPT (Improving Access to Psychological Therapies) contains data on long-term conditions, including chronic pain. IAPT linked to CRIS was available through the front-end interface of CRIS; however, I was not able to access or resolve the issues in order to access the front-end. Eventually, access to this was shut down for everyone due to the implementation of the national opt-out, and the inability to

	apply this i.e., remove patients who opted out, to the existing version of the interface. This meant there was no access to IAPT data, and there were plans to set up a new technology that applied opt-out and would be accessible through the ePJS CRIS interface. Due to time constraints and other commitments, I was not able to follow up on this.

Appendix 2 - HTN sample size simulation results

All classifiers, sample sizes, class proportions and score

Classifier	Sample Size	Class Proportion (class 1/ class 0)	F1-score Weighted avg. (95% CI)	AUC Score
Logistic Regression	200	99/01	1.00 (0.99-1)	0.43
		95/05	0.91 (0.68-1)	0.58
		90/10	0.91 (0.77-1)	0.33
		80/20	0.79 (0.64-0.95)	0.5
		70/30	0.65 (0.47-0.84)	0.38
		60/40	0.58 (0.39-0.77)	0.69
		50/50	0.53 (0.35-0.71)	0.66
	400	99/01	1.00 (0.99-1)	0.5
		95/05	0.96 (0.81-1)	0.96
		90/10	0.89 (0.79-0.98)	0.67
		80/20	0.73 (0.6-0.87)	0.65
		70/30	0.74 (0.6-0.87)	0.62
		60/40	0.61 (0.46-0.76)	0.56
		50/50	0.58 (0.46-0.7)	0.68
	500	99/01	0.98 (0.75-0.99)	0.96
		95/05	0.93 (0.9-0.97)	0.71

		90/10	0.93 (0.9-0.96)	0.5
		80/20	0.85 (0.76-0.93)	0.78
		70/30	0.76 (0.64-0.86)	0.66
		60/40	0.71 (0.58-0.81)	0.68
		50/50	0.69 (0.59-0.79)	0.77
	600	99/01	0.98 (0.74-0.99)	0.85
		95/05	0.92 (0.84-0.96)	0.55
		90/10	0.8 (0.7-0.89)	0.57
		80/20	0.81 (0.72-0.9)	0.71
		70/30	0.76 (0.65-0.86)	0.71
		60/40	0.64 (0.55-0.75)	0.68
		50/50	0.69 (0.59-0.78)	0.8
	800	99/01	1.00 (0.86-1)	0.81
		95/05	0.94 (0.83-0.96)	0.76
		90/10	0.92 (0.85-0.97)	0.65
		80/20	0.84 (0.76-0.91)	0.65
		70/30	0.77 (0.69-0.85)	0.71
		60/40	0.65 (0.57-0.73)	0.75
		50/50	0.69 (0.61-0.77)	0.72
	1000	99/01	0.99 (0.87-1)	0.57
		95/05	0.96 (0.85-0.97)	0.43
		90/10	0.82 (0.75-0.89)	0.72

		80/20	0.82 (0.74-0.89)	0.71
		70/30	0.7 (0.62-0.77)	0.62
		60/40	0.66 (0.58-0.74)	0.68
		50/50	0.67 (0.6-0.74)	0.74
	2000	99/01	0.99 (0.89-1)	0.68
		95/05	0.93 (0.91-0.96)	0.74
		90/10	0.88 (0.83-0.92)	0.66
		80/20	0.82 (0.78-0.87)	0.74
		70/30	0.76 (0.71-0.81)	0.75
		60/40	0.72 (0.67-0.78)	0.78
		50/50	0.67 (0.62-0.72)	0.74
	3000	99/01	0.98 (0.86-1)	0.66
		95/05	0.93 (0.88-0.96)	0.72
		90/10	0.91 (0.87-0.93)	0.73
		80/20	0.81 (0.77-0.85)	0.75
		70/30	0.77 (0.73-0.81)	0.77
		60/40	0.68 (0.63-0.73)	0.75
		50/50	0.69 (0.65-0.74)	0.76
	4000	99/01	0.99 (0.87-1)	0.7
		95/05	0.93 (0.88-0.96)	0.71
		90/10	0.89 (0.86-0.92)	0.7
		80/20	0.81 (0.77-0.84)	0.76

		70/30	0.77 (0.74-0.8)	0.79
		60/40	0.71 (0.68-0.75)	0.77
		50/50	0.7 (0.66-0.73)	0.78
	5000	99/01	0.98 (0.85-1)	0.72
		95/05	0.96 (0.89-0.99)	0.73
		90/10	0.91 (0.89-0.93)	0.73
		80/20	0.83 (0.86-0.8)	0.78
		70/30	0.78 (0.81-0.74)	0.74
		60/40	0.75 (0.78-0.72)	0.81
		50/50	0.73 (0.76-0.7)	0.82
Decision Tree	200	99/01	1 (0.89-1)	0.3
		95/05	0.91 (0.82-0.98)	0.39
		90/10	0.88 (0.74-0.98)	0.47
		80/20	0.82 (0.67-0.94)	0.65
		70/30	0.55 (0.39-0.74)	0.37
		60/40	0.45 (0.28-0.65)	0.47
		50/50	0.53 (0.35-0.69)	0.54
	400	99/01	1 (0.85-1)	0.49
		95/05	0.96 (0.82–0.99)	0.54
		90/10	0.89 (0.8-0.97)	0.58
		80/20	0.73 (0.62-0.84)	0.61
		70/30	0.74 (0.61-0.85)	0.64

		60/40	0.5 (0.38-0.63)	0.47
		50/50	0.49 (0.37-0.6)	0.49
	500	99/01	0.98 (0.75-1)	0.32
		95/05	0.89 (0.81-0.94)	0.44
		90/10	0.88 (0.83-0.92)	0.51
		80/20	0.72 (0.62-0.81)	0.59
		70/30	0.72 (0.6-0.81)	0.63
		60/40	0.66 (0.56-0.77)	0.64
		50/50	0.72 (0.63-0.81)	0.73
	600	99/01	0.98 (0.81-1)	0.37
		95/05	0.91 (0.78-0.99)	0.46
		90/10	0.78 (0.68-0.87)	0.49
		80/20	0.75 (0.65-0.84)	0.66
		70/30	0.77 (0.67-0.86)	0.71
		60/40	0.55 (0.45-0.64)	0.54
		50/50	0.66 (0.56-0.76)	0.67
	800	99/01	1 (0.80-1)	0.52
		95/05	0.95 (0.85-0.99)	0.64
		90/10	0.9 (0.84-0.95)	0.68
		80/20	0.8 (0.72-0.88)	0.62
		70/30	0.76 (0.68-0.83)	0.7
		60/40	0.63 (0.54-0.71)	0.62

		50/50	0.66 (0.58-0.73)	0.66
	1000	99/01	0.98 (0.87-1)	0.47
		95/05	0.95 (0.80-0.98)	0.51
		90/10	0.79 (0.86-0.72)	0.5
		80/20	0.76 (0.82-0.68)	0.62
		70/30	0.64 (0.72-0.56)	0.55
		60/40	0.66 (0.74-0.59)	0.64
		50/50	0.59 (0.68-0.51)	0.59
	2000	99/01	0.99 (0.82-1)	0.51
		95/05	0.91 (0.83-0.99)	0.58
		90/10	0.84 (0.88-0.8)	0.63
		80/20	0.77 (0.81-0.72)	0.67
		70/30	0.74 (0.79-0.68)	0.66
		60/40	0.64 (0.69-0.59)	0.63
		50/50	0.66 (0.71-0.61)	0.67
	3000	99/01	0.98 (0.91-0.99)	0.61
		95/05	0.93 (0.85-0.95)	0.66
		90/10	0.85 (0.88-0.82)	0.61
		80/20	0.74 (0.78-0.69)	0.6
		70/30	0.71 (0.75-0.67)	0.66
		60/40	0.64 (0.68-0.59)	0.62
		50/50	0.64 (0.67-0.59)	0.63

	4000	99/01	0.99 (0.89-1)	0.58
		95/05	0.93 (0.87-0.96)	0.6
		90/10	0.85 (0.87-0.82)	0.63
		80/20	0.76 (0.79-0.72)	0.64
		70/30	0.7 (0.74-0.67)	0.68
		60/40	0.67 (0.7-0.63)	0.67
		50/50	0.68 (0.71-0.65)	0.68
	5000	99/01	0.99 (0.89-1)	0.55
		95/05	0.94 (0.90-0.96)	0.61
		90/10	0.87 (0.9-0.85)	0.63
		80/20	0.78 (0.81-0.75)	0.65
		70/30	0.73 (0.76-0.69)	0.67
		60/40	0.72 (0.75-0.69)	0.71
		50/50	0.69 (0.72-0.66)	0.69
K-Nearest Neighbour	200	99/01	1 (0.90-1)	0.3
		95/05	0.91 (0.82-0.96)	0.34
		90/10	0.91 (0.77-1)	0.38
		80/20	0.84 (0.66-0.97)	0.54
		70/30	0.67 (0.51-0.84)	0.37
		60/40	0.54 (0.37-0.74)	0.62
		50/50	0.57 (0.39-0.72)	0.59
	400	99/01	1 (0.78-1)	0.45

		95/05	0.96 (0.91-1)	0.49
		90/10	0.9 (0.81-0.98)	0.53
		80/20	0.73 (0.61-0.86)	0.71
		70/30	0.68 (0.54-0.81)	0.45
		60/40	0.56 (0.44-0.67)	0.6
		50/50	0.65 (0.53-0.76)	0.65
	500	99/01	0.98 (0.90-1)	0.5
		95/05	0.93 (0.82-0.99)	0.58
		90/10	0.82 (0.72-0.91)	0.61
		80/20	0.82 (0.73-0.9)	0.76
		70/30	0.79 (0.69-0.87)	0.59
		60/40	0.6 (0.5-0.7)	0.64
		50/50	0.69 (0.59-0.78)	0.74
	600	99/01	0.98 (0.78-0.99)	0.56
		95/05	0.91 (0.78-0.95)	0.6
		90/10	0.82 (0.72-0.91)	0.63
		80/20	0.82 (0.73-0.9)	0.67
		70/30	0.79 (0.69-0.87)	0.73
		60/40	0.6 (0.5-0.7)	0.65
		50/50	0.69 (0.59-0.78)	0.74
	800	99/01	1 (0.80-1)	0.58
		95/05	0.95 (0.80-0.98)	0.6

		90/10	0.88 (0.81-0.93)	0.72
		80/20	0.82 (0.74-0.9)	0.61
		70/30	0.79 (0.71-0.87)	0.64
		60/40	0.64 (0.57-0.72)	0.74
		50/50	0.64 (0.55-0.73)	0.72
	1000	99/01	0.99 (0.80-1)	0.49
		95/05	0.95 (0.83-0.96)	0.53
		90/10	0.86 (0.79-0.92)	0.6
		80/20	0.81 (0.74-0.89)	0.64
		70/30	0.67 (0.58-0.75)	0.6
		60/40	0.59 (0.51-0.67)	0.61
		50/50	0.67 (0.6-0.74)	0.72
	2000	99/01	0.99 (0.79-1)	0.53
		95/05	0.93 (0.84-0.95)	0.59
		90/10	0.88 (0.83-0.92)	0.61
		80/20	0.82 (0.77-0.87)	0.65
		70/30	0.74 (0.68-0.79)	0.71
		60/40	0.72 (0.67-0.76)	0.78
		50/50	0.65 (0.6-0.7)	0.73
	3000	99/01	0.98 (0.83-1)	0.6
		95/05	0.95 (0.90-0.96)	0.68
		90/10	0.9 (0.87-0.93)	0.71

		80/20	0.77 (0.73-0.82)	0.69
		70/30	0.74 (0.7-0.78)	0.71
		60/40	0.68 (0.64-0.72)	0.71
		50/50	0.66 (0.61-0.7)	0.73
	4000	99/01	0.99 (0.85-1)	0.54
		95/05	0.93 (0.90-0.97)	0.61
		90/10	0.85 (0.82-0.88)	0.65
		80/20	0.81 (0.78-0.85)	0.71
		70/30	0.74 (0.7-0.77)	0.76
		60/40	0.66 (0.62-0.7)	0.73
		50/50	0.69 (0.66-0.73)	0.77
	5000	99/01	0.98 (0.88-0.99)	0.61
		95/05	0.96 (0.91-0.98)	0.68
		90/10	0.9 (0.87-0.93)	0.7
		80/20	0.81 (0.78-0.84)	0.72
		70/30	0.76 (0.72-0.79)	0.73
		60/40	0.71 (0.68-0.74)	0.77
		50/50	0.7 (0.67-0.73)	0.79
Linear Support Vector Classifier	200	99/01	1 (0.81-1)	0.2
		95/05	0.91 (0.80-0.99)	0.29
		90/10	0.91 (0.77-1)	0.30
		80/20	0.79 (0.6-0.94)	0.48

		70/30	0.64 (0.44-0.82)	0.35
		60/40	0.58 (0.4-0.78)	0.70
		50/50	0.49 (0.32-0.65)	0.66
	400	99/01	1 (0.88-1)	0.6
		95/05	0.96 (0.85-0.97)	0.65
		90/10	0.88 (0.77-0.98)	0.72
		80/20	0.73 (0.57-0.86)	0.64
		70/30	0.73 (0.59-0.85)	0.62
		60/40	0.63 (0.5-0.77)	0.57
		50/50	0.58 (0.47-0.71)	0.68
	500	99/01	0.98 (0.80-1)	0.4
		95/05	0.93 (0.85-0.95)	0.45
		90/10	0.93 (0.89-0.97)	0.52
		80/20	0.86 (0.77-0.94)	0.76
		70/30	0.76 (0.65-0.87)	0.66
		60/40	0.7 (0.58-0.8)	0.69
		50/50	0.65 (0.54-0.76)	0.76
	600	99/01	0.98 (0.84-1)	0.47
		95/05	0.92 (0.86-0.94)	0.52
		90/10	0.8 (0.69-0.89)	0.57
		80/20	0.82 (0.72-0.9)	0.69
		70/30	0.76 (0.64-0.86)	0.72

		60/40	0.63 (0.52-0.72)	0.68
		50/50	0.67 (0.57-0.76)	0.80
	800	99/01	1 (0.83-1)	0.56
		95/05	0.94 (0.81-0.96)	0.61
		90/10	0.92 (0.86-0.97)	0.65
		80/20	0.84 (0.75-0.9)	0.64
		70/30	0.78 (0.69-0.86)	0.72
		60/40	0.66 (0.57-0.75)	0.75
		50/50	0.65 (0.57-0.73)	0.71
	1000	99/01	0.99 (0.88-1)	0.64
		95/05	0.96 (0.89-0.99)	0.68
		90/10	0.84 (0.77-0.91)	0.73
		80/20	0.82 (0.75-0.88)	0.72
		70/30	0.7 (0.62-0.77)	0.62
		60/40	0.65 (0.58-0.72)	0.68
		50/50	0.66 (0.58-0.73)	0.73
	2000	99/01	0.99 (0.88-1)	0.53
		95/05	0.93 (0.87-0.95)	0.6
		90/10	0.88 (0.83-0.92)	0.63
		80/20	0.83 (0.78-0.88)	0.74
		70/30	0.76 (0.71-0.81)	0.75
		60/40	0.73 (0.68-0.79)	0.77

		50/50	0.66 (0.61-0.71)	0.73
	3000	99/01	0.98 (0.88-0.99)	0.58
		95/05	0.94 (0.86-0.99)	0.63
		90/10	0.91 (0.88-0.94)	0.72
		80/20	0.81 (0.77-0.85)	0.74
		70/30	0.78 (0.74-0.81)	0.77
		60/40	0.68 (0.64-0.72)	0.75
		50/50	0.69 (0.65-0.73)	0.76
	4000	99/01	0.99 (0.88-1)	0.59
		95/05	0.93 (0.89-0.95)	0.64
		90/10	0.89 (0.86-0.92)	0.70
		80/20	0.81 (0.77-0.85)	0.75
		70/30	0.78 (0.74-0.81)	0.78
		60/40	0.71 (0.67-0.75)	0.77
		50/50	0.69 (0.66-0.73)	0.78
	5000	99/01	0.98 (0.85-1)	0.64
		95/05	0.96 (0.92-0.99)	0.7
		90/10	0.91 (0.89-0.93)	0.72
		80/20	0.82 (0.79-0.85)	0.77
		70/30	0.77 (0.74-0.8)	0.74
		60/40	0.74 (0.71-0.77)	0.81
		50/50	0.73 (0.69-0.76)	0.81

Naive Bayes	200	99/01	1 (0.96-1)	0.55
		95/05	0.91 (0.87-0.94)	0.59
		90/10	0.91 (0.77-1)	0.63
		80/20	0.72 (0.52-0.91)	0.51
		70/30	0.73 (0.56-0.91)	0.51
		60/40	0.61 (0.41-0.79)	0.76
		50/50	0.62 (0.47-0.78)	0.60
	400	99/01	1 (0.96-1)	0.62
		95/05	0.93 (0.92-0.99)	0.68
		90/10	0.88 (0.77-0.95)	0.72
		80/20	0.73 (0.6-0.85)	0.54
		70/30	0.73 (0.58-0.86)	0.73
		60/40	0.62 (0.46-0.76)	0.54
		50/50	0.65 (0.52-0.76)	0.69
	500	99/01	0.98 (0.94-1)	0.5
		95/05	0.93 (0.89-0.96)	0.56
		90/10	0.94 (0.9-0.97)	0.60
		80/20	0.74 (0.62-0.84)	0.73
		70/30	0.73 (0.6-0.85)	0.71
		60/40	0.67 (0.54-0.8)	0.72
		50/50	0.68 (0.57-0.78)	0.75
	600	99/01	0.98 (0.94-1)	0.44

		95/05	0.92 (0.88-0.95)	0.48
		90/10	0.8 (0.71-0.9)	0.52
		80/20	0.69 (0.57-0.8)	0.68
		70/30	0.72 (0.61-0.83)	0.69
		60/40	0.62 (0.51-0.74)	0.67
		50/50	0.64 (0.55-0.75)	0.77
	800	99/01	1 (0.96-1)	0.58
		95/05	0.94 (0.9-0.97)	0.62
		90/10	0.86 (0.79-0.93)	0.66
		80/20	0.81 (0.72-0.89)	0.70
		70/30	0.77 (0.67-0.85)	0.70
		60/40	0.62 (0.52-0.72)	0.76
		50/50	0.72 (0.64-0.8)	0.74
	1000	99/01	0.99 (0.95-1)	0.66
		95/05	0.96 (0.92-0.99)	0.70
		90/10	0.83 (0.75-0.9)	0.74
		80/20	0.77 (0.69-0.85)	0.64
		70/30	0.71 (0.62-0.79)	0.61
		60/40	0.59 (0.5-0.68)	0.67
		50/50	0.71 (0.63-0.77)	0.72
	2000	99/01	0.99 (0.95-1)	0.61
		95/05	0.93 (0.89-0.96)	0.65

		90/10	0.85 (0.8-0.9)	0.69
		80/20	0.79 (0.74-0.85)	0.75
		70/30	0.76 (0.7-0.81)	0.74
		60/40	0.7 (0.63-0.76)	0.75
		50/50	0.62 (0.57-0.68)	0.71
	3000	99/01	0.98 (0.94-1)	0.62
		95/05	0.92 (0.89-0.96)	0.66
		90/10	0.86 (0.82-0.9)	0.70
		80/20	0.78 (0.73-0.82)	0.69
		70/30	0.72 (0.67-0.77)	0.71
		60/40	0.64 (0.58-0.69)	0.73
		50/50	0.63 (0.59-0.68)	0.71
	4000	99/01	0.99 (0.95-1)	0.59
		95/05	0.91 (0.89-0.96)	0.63
		90/10	0.84 (0.81-0.88)	0.67
		80/20	0.8 (0.76-0.85)	0.74
		70/30	0.73 (0.69-0.77)	0.74
		60/40	0.62 (0.57-0.66)	0.71
		50/50	0.66 (0.62-0.7)	0.74
	5000	99/01	0.98 (0.94-1)	0.59
		95/05	0.94 (0.92-0.99)	0.63
		90/10	0.89 (0.86-0.92)	0.67

		80/20	0.81 (0.78-0.85)	0.70
		70/30	0.75 (0.71-0.78)	0.74
		60/40	0.72 (0.69-0.76)	0.78
		50/50	0.69 (0.66-0.73)	0.78
Random Forest	200	99/01	1 (0.97-1)	0.48
		95/05	0.91 (0.88-0.96)	0.54
		90/10	0.91 (0.77-1)	0.60
		80/20	0.73 (0.56-0.91)	0.50
		70/30	0.71 (0.54-0.89)	0.31
		60/40	0.4 (0.2-0.61)	0.71
		50/50	0.35 (0.16-0.55)	0.57
	400	99/01	1 (0.97-1)	0.56
		95/05	0.96 (0.93-1)	0.62
		90/10	0.89 (0.79-0.98)	0.68
		80/20	0.73 (0.58-0.85)	0.61
		70/30	0.67 (0.53-0.79)	0.58
		60/40	0.56 (0.4-0.72)	0.56
		50/50	0.65 (0.51-0.77)	0.68
	500	99/01	0.98 (0.95-1)	0.31
		95/05	0.93 (0.9-0.97)	0.37
		90/10	0.93 (0.9-0.96)	0.43
		80/20	0.71 (0.58-0.82)	0.68

		70/30	0.71 (0.6-0.83)	0.65
		60/40	0.58 (0.46-0.72)	0.70
		50/50	0.63 (0.52-0.73)	0.72
	600	99/01	0.98 (0.95-1)	0.54
		95/05	0.92 (0.89-0.96)	0.60
		90/10	0.8 (0.7-0.89)	0.66
		80/20	0.69 (0.57-0.79)	0.71
		70/30	0.62 (0.5-0.75)	0.76
		60/40	0.57 (0.44-0.7)	0.64
		50/50	0.64 (0.53-0.74)	0.74
	800	99/01	1 (0.97-1)	0.57
		95/05	0.94 (0.91-0.98)	0.63
		90/10	0.86 (0.79-0.93)	0.69
		80/20	0.8 (0.7-0.88)	0.61
		70/30	0.71 (0.61-0.8)	0.69
		60/40	0.58 (0.47-0.68)	0.76
		50/50	0.7 (0.62-0.78)	0.76
	1000	99/01	0.99 (0.96-1)	0.56
		95/05	0.96 (0.93-1)	0.62
		90/10	0.83 (0.75-0.89)	0.68
		80/20	0.74 (0.66-0.82)	0.61
		70/30	0.66 (0.56-0.75)	0.63

		60/40	0.58 (0.48-0.67)	0.63
		50/50	0.68 (0.61-0.76)	0.70
	2000	99/01	0.99 (0.96-1)	0.53
		95/05	0.93 (0.9-0.97)	0.59
		90/10	0.84 (0.8-0.89)	0.65
		80/20	0.73 (0.67-0.79)	0.71
		70/30	0.72 (0.66-0.78)	0.72
		60/40	0.64 (0.58-0.7)	0.74
		50/50	0.61 (0.56-0.67)	0.71
	3000	99/01	0.98 (0.95-1)	0.59
		95/05	0.92 (0.89-0.96)	0.65
		90/10	0.86 (0.82-0.9)	0.71
		80/20	0.75 (0.7-0.79)	0.64
		70/30	0.66 (0.6-0.72)	0.71
		60/40	0.59 (0.54-0.64)	0.70
		50/50	0.61 (0.56-0.66)	0.67
	4000	99/01	0.99 (0.96-1)	0.56
		95/05	0.91 (0.88-0.95)	0.62
		90/10	0.84 (0.81-0.87)	0.68
		80/20	0.73 (0.69-0.77)	0.72
		70/30	0.65 (0.6-0.69)	0.70
		60/40	0.55 (0.5-0.6)	0.68

		50/50	0.64 (0.61-0.68)	0.70
	5000	99/01	0.98 (0.95-1)	0.60
		95/05	0.94 (0.91-0.98)	0.66
		90/10	0.87 (0.84-0.9)	0.72
		80/20	0.76 (0.72-0.8)	0.67
		70/30	0.67 (0.63-0.71)	0.71
		60/40	0.63 (0.59-0.67)	0.75
		50/50	0.66 (0.63-0.69)	0.75
Stochastic Gradient Descent	200	99/01	1 (0.95-1)	0.30
		95/05	0.91 (0.86-0.94)	0.34
		90/10	0.91 (0.77-1)	0.38
		80/20	0.79 (0.6-0.93)	0.52
		70/30	0.65 (0.47-0.81)	0.44
		60/40	0.62 (0.43-0.78)	0.69
		50/50	0.59 (0.42-0.74)	0.68
	400	99/01	1 (0.95-1)	0.62
		95/05	0.96 (0.91-0.99)	0.66
		90/10	0.87 (0.76-0.96)	0.70
		80/20	0.73 (0.59-0.86)	0.62
		70/30	0.73 (0.61-0.84)	0.64
		60/40	0.64 (0.52-0.77)	0.59
		50/50	0.62 (0.5-0.75)	0.66

	500	99/01	0.98 (0.93-1)	0.37
		95/05	0.91 (0.86-0.94)	0.41
		90/10	0.92 (0.88-0.95)	0.45
		80/20	0.83 (0.74-0.91)	0.71
		70/30	0.72 (0.62-0.82)	0.65
		60/40	0.67 (0.56-0.77)	0.69
		50/50	0.64 (0.52-0.75)	0.76
	600	99/01	0.98 (0.93-1)	0.50
		95/05	0.91 (0.86-0.94)	0.54
		90/10	0.8 (0.71-0.89)	0.58
		80/20	0.84 (0.74-0.92)	0.65
		70/30	0.74 (0.65-0.83)	0.72
		60/40	0.64 (0.53-0.74)	0.66
		50/50	0.64 (0.54-0.73)	0.76
	800	99/01	1 (0.95-1)	0.60
		95/05	0.94 (0.89-0.97)	0.64
		90/10	0.9 (0.84-0.96)	0.68
		80/20	0.8 (0.72-0.88)	0.63
		70/30	0.76 (0.68-0.84)	0.71
		60/40	0.65 (0.57-0.72)	0.72
		50/50	0.58 (0.5-0.67)	0.65
	1000	99/01	0.99 (0.94-1)	0.61

		95/05	0.95 (0.9-0.98)	0.65
		90/10	0.83 (0.76-0.9)	0.69
		80/20	0.83 (0.75-0.89)	0.72
		70/30	0.67 (0.59-0.75)	0.61
		60/40	0.67 (0.59-0.74)	0.71
		50/50	0.65 (0.58-0.72)	0.70
	2000	99/01	0.99 (0.94-1)	0.56
		95/05	0.92 (0.87-0.95)	0.60
		90/10	0.88 (0.83-0.92)	0.64
		80/20	0.83 (0.78-0.88)	0.74
		70/30	0.76 (0.72-0.81)	0.74
		60/40	0.72 (0.67-0.77)	0.75
		50/50	0.64 (0.59-0.69)	0.70
	3000	99/01	0.98 (0.93-1)	0.62
		95/05	0.93 (0.88-0.96)	0.66
		90/10	0.89 (0.85-0.92)	0.70
		80/20	0.8 (0.77-0.85)	0.74
		70/30	0.77 (0.73-0.81)	0.76
		60/40	0.68 (0.64-0.72)	0.73
		50/50	0.66 (0.62-0.7)	0.74
	4000	99/01	0.99 (0.94-1)	0.62
		95/05	0.92 (0.87-0.95)	0.66

		90/10	0.88 (0.85-0.91)	0.70
		80/20	0.79 (0.75-0.82)	0.73
		70/30	0.76 (0.72-0.79)	0.77
		60/40	0.7 (0.67-0.74)	0.76
		50/50	0.69 (0.66-0.72)	0.77
	5000	99/01	0.98 (0.93-1)	0.64
		95/05	0.94 (0.89-0.97)	0.68
		90/10	0.9 (0.87-0.92)	0.72
		80/20	0.82 (0.79-0.84)	0.76
		70/30	0.75 (0.72-0.78)	0.72
		60/40	0.74 (0.71-0.76)	0.80
		50/50	0.72 (0.69-0.75)	0.81
Support Vector Classifier	200	99/01	1 (0.93-1)	0.32
		95/05	0.91 (0.84-0.93)	0.38
		90/10	0.91 (0.77-1)	0.43
		80/20	0.73 (0.52-0.91)	0.49
		70/30	0.73 (0.56-0.91)	0.35
		60/40	0.56 (0.36-0.76)	0.78
		50/50	0.48 (0.3-0.68)	0.34
	400	99/01	1 (0.93-1)	0.67
		95/05	0.96 (0.89-0.98)	0.73
		90/10	0.89 (0.79-0.98)	0.78

		80/20	0.73 (0.58-0.86)	0.59
		70/30	0.73 (0.6-0.86)	0.67
		60/40	0.64 (0.47-0.79)	0.58
		50/50	0.65 (0.52-0.77)	0.69
	500	99/01	0.98 (0.91-1)	0.40
		95/05	0.93 (0.86-0.95)	0.46
		90/10	0.93 (0.9-0.96)	0.51
		80/20	0.79 (0.67-0.89)	0.75
		70/30	0.75 (0.63-0.86)	0.68
		60/40	0.69 (0.57-0.8)	0.73
		50/50	0.67 (0.56-0.77)	0.76
	600	99/01	0.98 (0.91-1)	0.46
		95/05	0.92 (0.85-0.94)	0.52
		90/10	0.8 (0.7-0.89)	0.57
		80/20	0.81 (0.69-0.9)	0.70
		70/30	0.76 (0.67-0.86)	0.73
		60/40	0.63 (0.51-0.74)	0.67
		50/50	0.63 (0.53-0.72)	0.78
	800	99/01	1 (0.93-1)	0.56
		95/05	0.94 (0.87-0.96)	0.62
		90/10	0.88 (0.81-0.94)	0.67
		80/20	0.83 (0.74-0.9)	0.66

		70/30	0.77 (0.67-0.85)	0.71
		60/40	0.64 (0.55-0.73)	0.78
		50/50	0.72 (0.64-0.79)	0.76
	1000	99/01	0.99 (0.92-1)	0.64
		95/05	0.96 (0.89-0.98)	0.70
		90/10	0.83 (0.76-0.9)	0.75
		80/20	0.79 (0.71-0.87)	0.71
		70/30	0.72 (0.64-0.8)	0.60
		60/40	0.59 (0.49-0.69)	0.66
		50/50	0.68 (0.6-0.75)	0.73
	2000	99/01	0.99 (0.92-1)	0.59
		95/05	0.93 (0.86-0.95)	0.65
		90/10	0.87 (0.82-0.92)	0.70
		80/20	0.82 (0.77-0.87)	0.77
		70/30	0.78 (0.72-0.83)	0.76
		60/40	0.7 (0.64-0.75)	0.77
		50/50	0.66 (0.6-0.71)	0.74
	3000	99/01	0.98 (0.91-1)	0.63
		95/05	0.93 (0.86-0.95)	0.69
		90/10	0.9 (0.86-0.93)	0.74
		80/20	0.79 (0.74-0.84)	0.74
		70/30	0.77 (0.72-0.81)	0.77

		60/40	0.69 (0.64-0.73)	0.75
		50/50	0.71 (0.67-0.76)	0.77
	4000	99/01	0.99 (0.92-1)	0.58
		95/05	0.92 (0.85-0.94)	0.64
		90/10	0.89 (0.86-0.92)	0.69
		80/20	0.81 (0.77-0.85)	0.75
		70/30	0.78 (0.74-0.81)	0.77
		60/40	0.68 (0.64-0.71)	0.78
		50/50	0.7 (0.67-0.74)	0.78
	5000	99/01	0.98 (0.91-1)	0.61
		95/05	0.96 (0.89-0.98)	0.67
		90/10	0.91 (0.88-0.93)	0.72
		80/20	0.83 (0.8-0.86)	0.78
		70/30	0.77 (0.74-0.81)	0.76
		60/40	0.76 (0.73-0.8)	0.82
		50/50	0.76 (0.73-0.79)	0.83
BERT_base	200	99/01	0.65 (0.45-0.74)	0.42
		95/05	0.78 (0.65-0.84)	0.46
		90/10	0.86 (0.71-0.96)	0.50
		80/20	0.59 (0.42-0.75)	0.50
		70/30	0.55 (0.39-0.75)	0.50
		60/40	0.81 (0.67-0.95)	0.77

		50/50	0.53 (0.36-0.68)	0.53
	400	99/01	0.69 (0.58-0.79)	0.42
		95/05	0.64 (0.54-0.73)	0.46
		90/10	0.8 (0.68-0.89)	0.50
		80/20	0.66 (0.53-0.8)	0.50
		70/30	0.53 (0.4-0.68)	0.50
		60/40	0.63 (0.51-0.74)	0.63
		50/50	0.65 (0.54-0.76)	0.68
	500	99/01	0.69 (0.6-0.78)	0.42
		95/05	0.9 (0.86-0.93)	0.46
		90/10	0.85 (0.77-0.93)	0.50
		80/20	0.75 (0.65-0.85)	0.60
		70/30	0.71 (0.61-0.81)	0.61
		60/40	0.64 (0.54-0.73)	0.65
		50/50	0.58 (0.48-0.66)	0.59
	600	99/01	0.67 (0.63-0.72)	0.42
		95/05	0.75 (0.66-0.83)	0.46
		90/10	0.84 (0.75-0.91)	0.50
		80/20	0.69 (0.58-0.79)	0.50
		70/30	0.62 (0.51-0.73)	0.56
		60/40	0.64 (0.55-0.72)	0.63
		50/50	0.62 (0.53-0.71)	0.62

	800	99/01	0.69 (0.66-0.72)	0.45
		95/05	0.63 (0.51-0.74)	0.49
		90/10	0.84 (0.77-0.91)	0.53
		80/20	0.75 (0.66-0.83)	0.62
		70/30	0.69 (0.6-0.78)	0.61
		60/40	0.59 (0.49-0.68)	0.59
		50/50	0.67 (0.59-0.74)	0.66
	1000	99/01	0.71 (0.61-0.81)	0.42
		95/05	0.64 (0.54-0.73)	0.46
		90/10	0.86 (0.8-0.91)	0.50
		80/20	0.74 (0.67-0.82)	0.50
		70/30	0.71 (0.64-0.77)	0.60
		60/40	0.61 (0.54-0.68)	0.58
		50/50	0.64 (0.58-0.71)	0.65
	2000	99/01	0.77 (0.74-0.8)	0.53
		95/05	0.53 (0.4-0.68)	0.57
		90/10	0.9 (0.86-0.93)	0.61
		80/20	0.8 (0.76-0.84)	0.65
		70/30	0.72 (0.66-0.77)	0.64
		60/40	0.67 (0.63-0.72)	0.65
		50/50	0.63 (0.58-0.68)	0.63
	3000	99/01	0.63 (0.51-0.74)	0.58

		95/05	0.69 (0.58-0.79)	0.62
		90/10	0.91 (0.88-0.94)	0.66
		80/20	0.8 (0.76-0.83)	0.62
		70/30	0.75 (0.71-0.79)	0.67
		60/40	0.67 (0.63-0.71)	0.66
		50/50	0.69 (0.65-0.73)	0.69
	4000	99/01	0.59 (0.49-0.68)	0.54
		95/05	0.86 (0.71-0.96)	0.58
		90/10	0.91 (0.88-0.93)	0.62
		80/20	0.84 (0.81-0.87)	0.66
		70/30	0.77 (0.74-0.8)	0.68
		60/40	0.7 (0.67-0.74)	0.67
		50/50	0.7 (0.67-0.73)	0.70
	5000	99/01	0.66 (0.53-0.8)	0.56
		95/05	0.63 (0.51-0.74)	0.60
		90/10	0.91 (0.89-0.93)	0.64
		80/20	0.83 (0.8-0.85)	0.66
		70/30	0.75 (0.72-0.77)	0.68
		60/40	0.69 (0.66-0.72)	0.67
		50/50	0.71 (0.68-0.73)	0.71

Appendix 3 - Diabetes sample size simulation results

All classifiers, sample sizes, class proportions and scores

Classifier	Sample Size	Class Proportion (class 1/ class 0)	F1-score Weighted avg. (95% CI)	AUC Score
Logistic Regression	200	99/01	1 (1-1)	0.52
		95/05	0.91 (0.77-1)	0.58
		90/10	0.82 (0.64-0.95)	0.71
		80/20	0.77 (0.6-0.91)	0.83
		70/30	0.83 (0.67-0.97)	0.85
		60/40	0.67 (0.48-0.83)	0.84
		50/50	0.75 (0.61-0.88)	0.78
	400	99/01	1 (1-1)	0.91
		95/05	0.96 (0.88-1)	0.97
		90/10	0.95 (0.88-1)	0.96
		80/20	0.87 (0.77-0.95)	0.87
		70/30	0.73 (0.61-0.85)	0.78
		60/40	0.69 (0.57-0.81)	0.75
		50/50	0.73 (0.61-0.84)	0.85
	500	99/01	0.98 (0.94-1)	0.96
		95/05	0.92 (0.85-0.98)	0.71

		90/10	0.84 (0.74-0.94)	0.71
		80/20	0.83 (0.73-0.92)	0.82
		70/30	0.74 (0.63-0.84)	0.85
		60/40	0.76 (0.66-0.85)	0.86
		50/50	0.71 (0.6-0.81)	0.82
	600	99/01	0.98 (0.95-1)	0.85
		95/05	0.92 (0.85-0.98)	0.56
		90/10	0.9 (0.82-0.97)	0.75
		80/20	0.75 (0.64-0.86)	0.71
		70/30	0.72 (0.61-0.81)	0.78
		60/40	0.72 (0.63-0.8)	0.80
		50/50	0.73 (0.64-0.82)	0.83
	800	99/01	1 (1-1)	0.70
		95/05	0.94 (0.88-0.99)	0.76
		90/10	0.9 (0.83-0.95)	0.86
		80/20	0.83 (0.75-0.91)	0.77
		70/30	0.75 (0.66-0.83)	0.79
		60/40	0.72 (0.65-0.79)	0.81
		50/50	0.69 (0.61-0.76)	0.78
	1000	99/01	0.99 (0.97-1)	0.57
		95/05	0.96 (0.93-0.99)	0.44
		90/10	0.88 (0.82-0.93)	0.84

		80/20	0.76 (0.69-0.84)	0.80
		70/30	0.79 (0.73-0.85)	0.85
		60/40	0.75 (0.67-0.81)	0.84
		50/50	0.74 (0.67-0.81)	0.84
	2000	99/01	0.99 (0.98-1)	0.68
		95/05	0.93 (0.89-0.96)	0.74
		90/10	0.89 (0.84-0.93)	0.80
		80/20	0.86 (0.82-0.9)	0.84
		70/30	0.79 (0.75-0.84)	0.86
		60/40	0.81 (0.76-0.85)	0.88
		50/50	0.76 (0.71-0.8)	0.86
	3000	99/01	0.98 (0.97-1)	0.58
		95/05	0.93 (0.9-0.96)	0.81
		90/10	0.92 (0.9-0.95)	0.88
		80/20	0.87 (0.83-0.9)	0.85
		70/30	0.83 (0.8-0.86)	0.88
		60/40	0.8 (0.77-0.84)	0.89
		50/50	0.8 (0.77-0.84)	0.88
	4000	99/01	0.99 (0.98-1)	0.71
		95/05	0.93 (0.9-0.95)	0.807
		90/10	0.93 (0.9-0.95)	0.863
		80/20	0.87 (0.84-0.9)	0.890

		70/30	0.86 (0.83-0.89)	0.909
		60/40	0.82 (0.79-0.85)	0.897
		50/50	0.8 (0.77-0.83)	0.893
	5000	99/01	0.98 (0.97-0.99)	0.751
		95/05	0.96 (0.94-0.97)	0.852
		90/10	0.92 (0.9-0.94)	0.866
		80/20	0.85 (0.83-0.88)	0.877
		70/30	0.86 (0.83-0.88)	0.899
		60/40	0.83 (0.81-0.86)	0.916
		50/50	0.85 (0.82-0.87)	0.918
Decision Tree	200	99/01	1 (1-1)	0.45
		95/05	0.91 (0.77-1)	0.50
		90/10	0.8 (0.63-0.95)	0.48
		80/20	0.82 (0.7-0.94)	0.76
		70/30	0.73 (0.58-0.86)	0.70
		60/40	0.69 (0.53-0.84)	0.69
		50/50	0.47 (0.31-0.66)	0.47
	400	99/01	1 (1-1)	0.62
		95/05	0.96 (0.88-1)	0.67
		90/10	0.93 (0.86-0.98)	0.48
		80/20	0.78 (0.68-0.87)	0.66
		70/30	0.68 (0.56-0.8)	0.53

		60/40	0.76 (0.64-0.86)	0.73
		50/50	0.71 (0.59-0.83)	0.71
	500	99/01	0.98 (0.94-1)	0.50
		95/05	0.9 (0.82-0.96)	0.47
		90/10	0.85 (0.76-0.93)	0.61
		80/20	0.83 (0.74-0.91)	0.69
		70/30	0.76 (0.66-0.85)	0.73
		60/40	0.78 (0.69-0.86)	0.78
		50/50	0.84 (0.75-0.91)	0.84
	600	99/01	0.98 (0.95-1)	0.50
		95/05	0.91 (0.83-0.97)	0.48
		90/10	0.89 (0.82-0.96)	0.64
		80/20	0.68 (0.58-0.79)	0.51
		70/30	0.66 (0.56-0.76)	0.61
		60/40	0.63 (0.53-0.71)	0.64
		50/50	0.7 (0.61-0.78)	0.71
	800	99/01	1 (1-1)	0.54
		95/05	0.95 (0.9-0.99)	0.59
		90/10	0.89 (0.84-0.95)	0.72
		80/20	0.78 (0.7-0.85)	0.61
		70/30	0.75 (0.66-0.82)	0.69
		60/40	0.67 (0.59-0.76)	0.66

		50/50	0.68 (0.6-0.75)	0.69
	1000	99/01	0.98 (0.96-1)	0.49
		95/05	0.95 (0.92-0.98)	0.60
		90/10	0.86 (0.81-0.91)	0.69
		80/20	0.72 (0.64-0.79)	0.60
		70/30	0.8 (0.73-0.86)	0.73
		60/40	0.7 (0.63-0.76)	0.70
		50/50	0.61 (0.55-0.69)	0.61
	2000	99/01	0.99 (0.97-1)	0.49
		95/05	0.91 (0.88-0.95)	0.51
		90/10	0.88 (0.84-0.91)	0.60
		80/20	0.85 (0.81-0.88)	0.75
		70/30	0.78 (0.74-0.83)	0.74
		60/40	0.8 (0.75-0.83)	0.80
		50/50	0.76 (0.72-0.81)	0.77
	3000	99/01	0.98 (0.96-0.99)	0.50
		95/05	0.93 (0.91-0.96)	0.69
		90/10	0.92 (0.89-0.95)	0.79
		80/20	0.83 (0.8-0.86)	0.76
		70/30	0.77 (0.74-0.8)	0.74
		60/40	0.82 (0.78-0.85)	0.83
		50/50	0.8 (0.76-0.83)	0.80

	4000	99/01	0.99 (0.98-0.99)	0.50
		95/05	0.93 (0.91-0.95)	0.65
		90/10	0.9 (0.88-0.92)	0.75
		80/20	0.86 (0.83-0.89)	0.79
		70/30	0.85 (0.82-0.87)	0.84
		60/40	0.83 (0.8-0.86)	0.83
		50/50	0.82 (0.79-0.85)	0.82
	5000	99/01	0.98 (0.97-0.99)	0.65
		95/05	0.94 (0.93-0.96)	0.68
		90/10	0.9 (0.88-0.92)	0.76
		80/20	0.87 (0.84-0.89)	0.80
		70/30	0.87 (0.84-0.89)	0.86
		60/40	0.85 (0.83-0.87)	0.85
		50/50	0.87 (0.85-0.9)	0.87
K-Nearest Neighbour	200	99/01	1 (1-1)	0.61
		95/05	0.9 (0.77-1)	0.67
		90/10	0.82 (0.6-0.95)	0.45
		80/20	0.86 (0.71-0.97)	0.81
		70/30	0.81 (0.66-0.94)	0.82
		60/40	0.69 (0.53-0.84)	0.76
		50/50	0.58 (0.41-0.75)	0.71
	400	99/01	1 (1-1)	0.90

		95/05	0.96 (0.9-1)	0.96
		90/10	0.97 (0.92-1)	0.94
		80/20	0.86 (0.78-0.94)	0.88
		70/30	0.69 (0.57-0.8)	0.73
		60/40	0.71 (0.59-0.82)	0.72
		50/50	0.68 (0.56-0.8)	0.79
	500	99/01	0.98 (0.94-1)	0.49
		95/05	0.93 (0.85-0.99)	0.51
		90/10	0.84 (0.75-0.92)	0.67
		80/20	0.87 (0.79-0.94)	0.76
		70/30	0.72 (0.6-0.82)	0.81
		60/40	0.73 (0.63-0.83)	0.81
		50/50	0.69 (0.59-0.79)	0.77
	600	99/01	0.98 (0.95-1)	0.47
		95/05	0.91 (0.84-0.97)	0.52
		90/10	0.89 (0.81-0.96)	0.65
		80/20	0.76 (0.66-0.84)	0.74
		70/30	0.77 (0.68-0.86)	0.75
		60/40	0.68 (0.59-0.78)	0.75
		50/50	0.68 (0.58-0.77)	0.77
	800	99/01	1 (1-1)	0.55
		95/05	0.95 (0.9-0.99)	0.61

		90/10	0.92 (0.85-0.97)	0.79
		80/20	0.81 (0.74-0.88)	0.78
		70/30	0.7 (0.61-0.79)	0.74
		60/40	0.71 (0.64-0.78)	0.80
		50/50	0.62 (0.53-0.7)	0.70
	1000	99/01	0.99 (0.97-1)	0.46
		95/05	0.95 (0.92-0.98)	0.54
		90/10	0.87 (0.81-0.92)	0.73
		80/20	0.76 (0.68-0.84)	0.74
		70/30	0.81 (0.75-0.87)	0.83
		60/40	0.77 (0.7-0.83)	0.82
		50/50	0.7 (0.63-0.77)	0.77
	2000	99/01	0.99 (0.98-1)	0.48
		95/05	0.93 (0.9-0.96)	0.57
		90/10	0.88 (0.84-0.92)	0.68
		80/20	0.79 (0.74-0.84)	0.73
		70/30	0.74 (0.68-0.79)	0.81
		60/40	0.73 (0.68-0.77)	0.81
		50/50	0.71 (0.66-0.77)	0.79
	3000	99/01	0.98 (0.96-0.99)	0.57
		95/05	0.95 (0.92-0.97)	0.73
		90/10	0.9 (0.87-0.92)	0.80

		80/20	0.81 (0.77-0.85)	0.82
		70/30	0.77 (0.73-0.8)	0.81
		60/40	0.77 (0.73-0.81)	0.85
		50/50	0.73 (0.69-0.77)	0.83
	4000	99/01	0.99 (0.98-1)	0.47
		95/05	0.93 (0.91-0.95)	0.65
		90/10	0.92 (0.89-0.94)	0.77
		80/20	0.82 (0.79-0.85)	0.82
		70/30	0.78 (0.75-0.82)	0.82
		60/40	0.77 (0.74-0.8)	0.85
		50/50	0.76 (0.73-0.8)	0.84
	5000	99/01	0.98 (0.97-0.99)	0.69
		95/05	0.96 (0.94-0.97)	0.71
		90/10	0.91 (0.88-0.93)	0.80
		80/20	0.85 (0.82-0.88)	0.82
		70/30	0.8 (0.77-0.83)	0.85
		60/40	0.78 (0.75-0.8)	0.87
		50/50	0.76 (0.73-0.79)	0.85
Linear Support Vector Classifier	200	99/01	1 (1-1)	0.11
		95/05	0.91 (0.77-1)	0.13
		90/10	0.82 (0.64-0.95)	0.67
		80/20	0.81 (0.67-0.94)	0.81

		70/30	0.84 (0.69-0.97)	0.83
		60/40	0.65 (0.47-0.81)	0.84
		50/50	0.75 (0.61-0.91)	0.78
	400	99/01	1 (1-1)	0.94
		95/05	0.96 (0.9-1)	0.96
		90/10	0.95 (0.88-1)	0.96
		80/20	0.92 (0.84-0.98)	0.88
		70/30	0.71 (0.58-0.84)	0.79
		60/40	0.66 (0.55-0.77)	0.75
		50/50	0.71 (0.58-0.82)	0.85
	500	99/01	0.98 (0.94-1)	0.71
		95/05	0.92 (0.85-0.98)	0.73
		90/10	0.87 (0.78-0.95)	0.69
		80/20	0.85 (0.76-0.93)	0.79
		70/30	0.74 (0.64-0.84)	0.86
		60/40	0.76 (0.66-0.85)	0.86
		50/50	0.74 (0.64-0.83)	0.82
	600	99/01	0.98 (0.95-1)	0.97
		95/05	0.92 (0.84-0.98)	0.41
		90/10	0.91 (0.82-0.97)	0.69
		80/20	0.75 (0.63-0.85)	0.70
		70/30	0.71 (0.61-0.8)	0.78

		60/40	0.7 (0.6-0.78)	0.79
		50/50	0.73 (0.64-0.8)	0.82
	800	99/01	1 (1-1)	0.68
		95/05	0.94 (0.9-0.99)	0.70
		90/10	0.89 (0.82-0.95)	0.89
		80/20	0.86 (0.78-0.92)	0.73
		70/30	0.75 (0.65-0.83)	0.78
		60/40	0.72 (0.63-0.79)	0.80
		50/50	0.7 (0.62-0.78)	0.78
	1000	99/01	0.99 (0.97-1)	0.49
		95/05	0.96 (0.92-0.99)	0.35
		90/10	0.88 (0.82-0.94)	0.80
		80/20	0.76 (0.67-0.83)	0.77
		70/30	0.81 (0.74-0.87)	0.84
		60/40	0.76 (0.7-0.83)	0.84
		50/50	0.74 (0.66-0.81)	0.85
	2000	99/01	0.99 (0.98-1)	0.44
		95/05	0.93 (0.9-0.96)	0.74
		90/10	0.89 (0.84-0.93)	0.75
		80/20	0.86 (0.82-0.9)	0.83
		70/30	0.81 (0.76-0.85)	0.85
		60/40	0.81 (0.76-0.85)	0.88

		50/50	0.76 (0.72-0.81)	0.87
	3000	99/01	0.98 (0.97-0.99)	0.46
		95/05	0.94 (0.91-0.97)	0.77
		90/10	0.93 (0.9-0.96)	0.86
		80/20	0.86 (0.82-0.89)	0.84
		70/30	0.83 (0.8-0.86)	0.87
		60/40	0.8 (0.77-0.84)	0.89
		50/50	0.8 (0.77-0.84)	0.88
	4000	99/01	0.99 (0.98-1)	0.60
		95/05	0.93 (0.91-0.96)	0.76
		90/10	0.93 (0.91-0.96)	0.85
		80/20	0.87 (0.84-0.9)	0.88
		70/30	0.86 (0.83-0.88)	0.91
		60/40	0.83 (0.8-0.86)	0.90
		50/50	0.81 (0.78-0.84)	0.89
	5000	99/01	0.98 (0.97-0.99)	0.73
		95/05	0.96 (0.94-0.97)	0.81
		90/10	0.93 (0.9-0.95)	0.84
		80/20	0.86 (0.83-0.88)	0.87
		70/30	0.86 (0.83-0.88)	0.89
		60/40	0.82 (0.8-0.85)	0.91
		50/50	0.83 (0.81-0.86)	0.92

Naive Bayes	200	99/01	1 (1-1)	0.28
		95/05	0.91 (0.77-1)	0.33
		90/10	0.82 (0.64-0.95)	0.79
		80/20	0.73 (0.56-0.91)	0.81
		70/30	0.67 (0.46-0.85)	0.80
		60/40	0.48 (0.27-0.67)	0.82
		50/50	0.66 (0.48-0.81)	0.78
	400	99/01	1 (1-1)	0.46
		95/05	0.93 (0.84-1)	0.51
		90/10	0.95 (0.88-1)	0.91
		80/20	0.83 (0.71-0.94)	0.84
		70/30	0.61 (0.49-0.75)	0.75
		60/40	0.73 (0.62-0.86)	0.81
		50/50	0.62 (0.5-0.74)	0.79
	500	99/01	0.98 (0.94-1)	0.57
		95/05	0.93 (0.85-0.98)	0.50
		90/10	0.82 (0.7-0.93)	0.63
		80/20	0.75 (0.63-0.85)	0.77
		70/30	0.62 (0.49-0.75)	0.83
		60/40	0.72 (0.61-0.82)	0.83
		50/50	0.64 (0.53-0.75)	0.77
	600	99/01	0.98 (0.95-1)	0.88

		95/05	0.92 (0.85-0.98)	0.74
		90/10	0.86 (0.76-0.94)	0.53
		80/20	0.71 (0.59-0.83)	0.69
		70/30	0.62 (0.48-0.74)	0.75
		60/40	0.69 (0.57-0.78)	0.80
		50/50	0.73 (0.64-0.81)	0.84
	800	99/01	1 (1-1)	0.65
		95/05	0.94 (0.88-0.99)	0.70
		90/10	0.86 (0.78-0.93)	0.85
		80/20	0.74 (0.64-0.83)	0.73
		70/30	0.62 (0.51-0.72)	0.75
		60/40	0.68 (0.59-0.77)	0.84
		50/50	0.68 (0.6-0.77)	0.77
	1000	99/01	0.99 (0.97-1)	0.97
		95/05	0.96 (0.93-0.99)	0.43
		90/10	0.83 (0.75-0.9)	0.70
		80/20	0.71 (0.62-0.79)	0.68
		70/30	0.73 (0.64-0.82)	0.79
		60/40	0.7 (0.62-0.77)	0.79
		50/50	0.69 (0.6-0.77)	0.76
	2000	99/01	0.99 (0.98-1)	0.57
		95/05	0.93 (0.89-0.96)	0.58

		90/10	0.87 (0.82-0.91)	0.69
		80/20	0.76 (0.69-0.81)	0.76
		70/30	0.67 (0.61-0.74)	0.77
		60/40	0.69 (0.63-0.75)	0.83
		50/50	0.69 (0.64-0.74)	0.79
	3000	99/01	0.98 (0.97-0.99)	0.48
		95/05	0.92 (0.89-0.94)	0.71
		90/10	0.87 (0.84-0.91)	0.76
		80/20	0.77 (0.72-0.82)	0.79
		70/30	0.76 (0.71-0.81)	0.85
		60/40	0.74 (0.7-0.78)	0.83
		50/50	0.74 (0.7-0.77)	0.82
	4000	99/01	0.99 (0.98-1)	0.49
		95/05	0.91 (0.88-0.94)	0.66
		90/10	0.9 (0.87-0.93)	0.74
		80/20	0.76 (0.72-0.8)	0.78
		70/30	0.73 (0.69-0.77)	0.82
		60/40	0.68 (0.64-0.72)	0.83
		50/50	0.72 (0.68-0.75)	0.82
	5000	99/01	0.98 (0.97-0.99)	0.62
		95/05	0.94 (0.92-0.96)	0.73
		90/10	0.88 (0.85-0.91)	0.73

		80/20	0.8 (0.77-0.84)	0.80
		70/30	0.72 (0.68-0.76)	0.83
		60/40	0.7 (0.66-0.73)	0.85
		50/50	0.76 (0.73-0.79)	0.86
Random Forest	200	99/01	1 (1-1)	0.49
		95/05	0.91 (0.77-1)	0.55
		90/10	0.81 (0.69-0.95)	0.63
		80/20	0.74 (0.54-0.95)	0.78
		70/30	0.77 (0.59-0.93)	0.89
		60/40	0.48 (0.27-0.71)	0.75
		50/50	0.75 (0.6-0.88)	0.87
	400	99/01	1 (1-1)	0.91
		95/05	0.96 (0.89-1)	0.97
		90/10	0.95 (0.88-1)	0.87
		80/20	0.84 (0.73-0.95)	0.74
		70/30	0.62 (0.48-0.75)	0.76
		60/40	0.71 (0.59-0.83)	0.76
		50/50	0.63 (0.49-0.76)	0.76
	500	99/01	0.98 (0.94-1)	0.82
		95/05	0.93 (0.85-0.98)	0.73
		90/10	0.82 (0.71-0.91)	0.79
		80/20	0.75 (0.63-0.85)	0.83

		70/30	0.5 (0.38-0.63)	0.76
		60/40	0.72 (0.6-0.82)	0.83
		50/50	0.65 (0.55-0.76)	0.77
	600	99/01	0.99 (0.95-1)	0.99
		95/05	0.92 (0.85-0.98)	0.49
		90/10	0.86 (0.77-0.94)	0.77
		80/20	0.68 (0.57-0.8)	0.78
		70/30	0.62 (0.5-0.74)	0.69
		60/40	0.61 (0.5-0.72)	0.76
		50/50	0.61 (0.51-0.71)	0.81
	800	99/01	1 (1-1)	0.61
		95/05	0.94 (0.88-0.98)	0.67
		90/10	0.86 (0.78-0.93)	0.82
		80/20	0.76 (0.65-0.84)	0.80
		70/30	0.62 (0.51-0.72)	0.76
		60/40	0.65 (0.56-0.75)	0.84
		50/50	0.67 (0.58-0.76)	0.77
	1000	99/01	0.99 (0.97-1)	0.52
		95/05	0.96 (0.93-0.99)	0.61
		90/10	0.84 (0.77-0.91)	0.80
		80/20	0.68 (0.6-0.77)	0.78
		70/30	0.7 (0.61-0.79)	0.87

		60/40	0.6 (0.5-0.69)	0.81
		50/50	0.6 (0.52-0.69)	0.70
	2000	99/01	0.99 (0.98-1)	0.67
		95/05	0.93 (0.89-0.96)	0.58
		90/10	0.87 (0.82-0.91)	0.70
		80/20	0.73 (0.67-0.8)	0.76
		70/30	0.65 (0.59-0.71)	0.80
		60/40	0.61 (0.54-0.66)	0.75
		50/50	0.65 (0.6-0.7)	0.74
	3000	99/01	0.98 (0.97-1)	0.48
		95/05	0.92 (0.9-0.95)	0.74
		90/10	0.88 (0.84-0.91)	0.76
		80/20	0.75 (0.7-0.79)	0.79
		70/30	0.67 (0.61-0.72)	0.80
		60/40	0.66 (0.61-0.71)	0.78
		50/50	0.72 (0.67-0.76)	0.81
	4000	99/01	0.99 (0.98-1)	0.53
		95/05	0.91 (0.88-0.94)	0.70
		90/10	0.88 (0.85-0.91)	0.76
		80/20	0.74 (0.69-0.78)	0.76
		70/30	0.67 (0.63-0.72)	0.81
		60/40	0.6 (0.56-0.64)	0.79

		50/50	0.67 (0.63-0.71)	0.77
	5000	99/01	0.98 (0.97-0.99)	0.75
		95/05	0.94 (0.92-0.96)	0.76
		90/10	0.86 (0.83-0.89)	0.84
		80/20	0.74 (0.71-0.79)	0.80
		70/30	0.64 (0.6-0.68)	0.81
		60/40	0.64 (0.6-0.69)	0.79
		50/50	0.71 (0.68-0.75)	0.81
Stochastic Gradient Descent	200	99/01	1 (1-1)	0.55
		95/05	0.91 (0.77-1)	0.62
		90/10	0.81 (0.64-0.95)	0.69
		80/20	0.85 (0.7-0.95)	0.82
		70/30	0.77 (0.62-0.91)	0.78
		60/40	0.67 (0.49-0.84)	0.76
		50/50	0.69 (0.53-0.85)	0.73
	400	99/01	1 (1-1)	0.89
		95/05	0.96 (0.88-1)	0.96
		90/10	0.97 (0.92-1)	0.97
		80/20	0.88 (0.8-0.95)	0.90
		70/30	0.67 (0.53-0.79)	0.78
		60/40	0.68 (0.56-0.79)	0.71
		50/50	0.72 (0.6-0.83)	0.83

	500	99/01	0.98 (0.94-1)	0.37
		95/05	0.91 (0.84-0.98)	0.67
		90/10	0.88 (0.79-0.96)	0.60
		80/20	0.84 (0.75-0.92)	0.74
		70/30	0.78 (0.68-0.87)	0.86
		60/40	0.8 (0.7-0.89)	0.81
		50/50	0.74 (0.65-0.83)	0.78
	600	99/01	0.98 (0.95-1)	0.99
		95/05	0.91 (0.84-0.97)	0.39
		90/10	0.91 (0.84-0.97)	0.65
		80/20	0.74 (0.63-0.84)	0.64
		70/30	0.72 (0.63-0.82)	0.73
		60/40	0.62 (0.52-0.72)	0.73
		50/50	0.72 (0.63-0.81)	0.79
	800	99/01	1 (1-1)	0.54
		95/05	0.94 (0.88-0.98)	0.61
		90/10	0.9 (0.84-0.96)	0.88
		80/20	0.84 (0.75-0.9)	0.67
		70/30	0.78 (0.7-0.85)	0.74
		60/40	0.7 (0.62-0.78)	0.75
		50/50	0.68 (0.6-0.76)	0.73
	1000	99/01	0.99 (0.97-1)	0.75

		95/05	0.95 (0.92-0.99)	0.21
		90/10	0.89 (0.84-0.94)	0.78
		80/20	0.75 (0.68-0.82)	0.70
		70/30	0.78 (0.7-0.84)	0.81
		60/40	0.78 (0.71-0.85)	0.81
		50/50	0.72 (0.65-0.79)	0.82
	2000	99/01	0.99 (0.98-1)	0.42
		95/05	0.92 (0.88-0.95)	0.76
		90/10	0.89 (0.84-0.92)	0.69
		80/20	0.87 (0.83-0.91)	0.83
		70/30	0.79 (0.75-0.84)	0.84
		60/40	0.8 (0.76-0.84)	0.86
		50/50	0.78 (0.73-0.82)	0.86
	3000	99/01	0.98 (0.96-0.99)	0.33
		95/05	0.93 (0.9-0.95)	0.76
		90/10	0.92 (0.9-0.95)	0.85
		80/20	0.87 (0.84-0.9)	0.82
		70/30	0.81 (0.78-0.85)	0.85
		60/40	0.81 (0.77-0.84)	0.88
		50/50	0.83 (0.8-0.86)	0.87
	4000	99/01	0.99 (0.98-1)	0.60
		95/05	0.92 (0.89-0.94)	0.77

		90/10	0.93 (0.9-0.95)	0.83
		80/20	0.88 (0.85-0.9)	0.89
		70/30	0.86 (0.83-0.88)	0.90
		60/40	0.82 (0.79-0.85)	0.89
		50/50	0.8 (0.76-0.83)	0.88
	5000	99/01	0.98 (0.97-0.99)	0.72
		95/05	0.94 (0.93-0.96)	0.81
		90/10	0.9 (0.88-0.93)	0.85
		80/20	0.88 (0.85-0.9)	0.87
		70/30	0.84 (0.81-0.87)	0.89
		60/40	0.82 (0.79-0.85)	0.90
		50/50	0.84 (0.82-0.87)	0.90
Support Vector Classifier	200	99/01	1 (1-1)	0.64
		95/05	0.9 (0.77-1)	0.68
		90/10	0.82 (0.64-0.98)	0.65
		80/20	0.79 (0.6-0.93)	0.78
		70/30	0.77 (0.59-0.93)	0.82
		60/40	0.37 (0.18-0.57)	0.84
		50/50	0.55 (0.36-0.72)	0.79
	400	99/01	1 (1-1)	0.93
		95/05	0.96 (0.9-1)	0.97
		90/10	0.95 (0.88-1)	0.95

		80/20	0.84 (0.72-0.94)	0.90
		70/30	0.69 (0.54-0.82)	0.79
		60/40	0.74 (0.61-0.86)	0.76
		50/50	0.61 (0.48-0.73)	0.84
	500	99/01	0.98 (0.94-1)	0.22
		95/05	0.92 (0.85-0.98)	0.29
		90/10	0.82 (0.71-0.91)	0.70
		80/20	0.77 (0.65-0.88)	0.80
		70/30	0.68 (0.56-0.8)	0.85
		60/40	0.74 (0.63-0.83)	0.83
		50/50	0.71 (0.61-0.81)	0.79
	600	99/01	0.98 (0.95-1)	0.02
		95/05	0.93 (0.86-0.98)	0.44
		90/10	0.86 (0.77-0.94)	0.64
		80/20	0.73 (0.61-0.85)	0.76
		70/30	0.71 (0.6-0.8)	0.78
		60/40	0.7 (0.6-0.8)	0.77
		50/50	0.73 (0.65-0.82)	0.82
	800	99/01	1 (1-1)	0.57
		95/05	0.94 (0.88-0.99)	0.61
		90/10	0.86 (0.78-0.93)	0.89
		80/20	0.81 (0.72-0.89)	0.77

		70/30	0.75 (0.66-0.83)	0.77
		60/40	0.73 (0.66-0.81)	0.79
		50/50	0.71 (0.63-0.78)	0.76
	1000	99/01	0.99 (0.97-1)	0.52
		95/05	0.96 (0.93-0.99)	0.38
		90/10	0.84 (0.77-0.91)	0.80
		80/20	0.74 (0.66-0.82)	0.74
		70/30	0.81 (0.74-0.87)	0.80
		60/40	0.75 (0.68-0.82)	0.84
		50/50	0.75 (0.68-0.81)	0.83
	2000	99/01	0.99 (0.97-1)	0.43
		95/05	0.93 (0.89-0.96)	0.75
		90/10	0.89 (0.84-0.93)	0.71
		80/20	0.82 (0.77-0.87)	0.83
		70/30	0.82 (0.77-0.87)	0.86
		60/40	0.83 (0.78-0.87)	0.89
		50/50	0.78 (0.72-0.82)	0.86
	3000	99/01	0.98 (0.96-0.99)	0.39
		95/05	0.93 (0.9-0.96)	0.77
		90/10	0.93 (0.9-0.95)	0.83
		80/20	0.86 (0.82-0.89)	0.83
		70/30	0.84 (0.8-0.88)	0.86

		60/40	0.84 (0.81-0.87)	0.90
		50/50	0.83 (0.79-0.86)	0.90
	4000	99/01	0.99 (0.98-1)	0.42
		95/05	0.92 (0.89-0.95)	0.69
		90/10	0.92 (0.89-0.94)	0.85
		80/20	0.86 (0.82-0.89)	0.87
		70/30	0.86 (0.83-0.89)	0.90
		60/40	0.86 (0.84-0.89)	0.92
		50/50	0.83 (0.8-0.85)	0.91
	5000	99/01	0.98 (0.97-0.99)	0.66
		95/05	0.96 (0.94-0.98)	0.81
		90/10	0.92 (0.89-0.94)	0.82
		80/20	0.86 (0.84-0.89)	0.85
		70/30	0.87 (0.85-0.89)	0.91
		60/40	0.86 (0.84-0.88)	0.92
		50/50	0.86 (0.84-0.89)	0.93
BERT_base	200	99/01	1 (1-1)	0.44
		95/05	0.93 (0.82-1)	0.50
		90/10	0.85 (0.71-0.96)	0.50
		80/20	0.63 (0.45-0.78)	0.50
		70/30	0.45 (0.28-0.61)	0.50
		60/40	0.61 (0.44-0.76)	0.59

		50/50	0.6 (0.47-0.75)	0.60
	400	99/01	1 (1-1)	0.44
		95/05	0.98 (0.94-1)	0.50
		90/10	0.82 (0.71-0.91)	0.50
		80/20	0.61 (0.5-0.73)	0.50
		70/30	0.76 (0.65-0.86)	0.64
		60/40	0.68 (0.58-0.78)	0.67
		50/50	0.54 (0.41-0.67)	0.58
	500	99/01	0.98 (0.94-1)	0.50
		95/05	0.97 (0.93-1)	0.50
		90/10	0.72 (0.62-0.83)	0.50
		80/20	0.79 (0.69-0.88)	0.59
		70/30	0.57 (0.46-0.69)	0.52
		60/40	0.62 (0.52-0.71)	0.62
		50/50	0.65 (0.56-0.74)	0.67
	600	99/01	0.98 (0.94-1)	0.50
		95/05	0.91 (0.85-0.98)	0.50
		90/10	0.81 (0.73-0.89)	0.50
		80/20	0.65 (0.54-0.77)	0.52
		70/30	0.51 (0.4-0.62)	0.51
		60/40	0.67 (0.58-0.76)	0.67
		50/50	0.7 (0.63-0.78)	0.70

	800	99/01	0.98 (0.95-1)	0.50
		95/05	0.92 (0.86-0.96)	0.50
		90/10	0.82 (0.75-0.89)	0.50
		80/20	0.69 (0.61-0.79)	0.57
		70/30	0.77 (0.71-0.84)	0.69
		60/40	0.72 (0.65-0.79)	0.72
		50/50	0.72 (0.65-0.8)	0.73
	1000	99/01	0.98 (0.95-1)	0.50
		95/05	0.96 (0.93-0.99)	0.50
		90/10	0.86 (0.8-0.91)	0.50
		80/20	0.74 (0.66-0.81)	0.50
		70/30	0.79 (0.74-0.85)	0.73
		60/40	0.74 (0.68-0.8)	0.73
		50/50	0.73 (0.67-0.78)	0.72
	2000	99/01	0.98 (0.96-0.99)	0.50
		95/05	0.97 (0.95-0.99)	0.63
		90/10	0.87 (0.83-0.91)	0.59
		80/20	0.83 (0.79-0.87)	0.73
		70/30	0.78 (0.74-0.82)	0.72
		60/40	0.75 (0.7-0.79)	0.74
		50/50	0.78 (0.73-0.82)	0.78
	3000	99/01	0.98 (0.97-0.99)	0.50

		95/05	0.92 (0.9-0.95)	0.59
		90/10	0.89 (0.85-0.91)	0.58
		80/20	0.83 (0.8-0.87)	0.70
		70/30	0.8 (0.76-0.83)	0.75
		60/40	0.77 (0.73-0.8)	0.76
		50/50	0.76 (0.73-0.8)	0.76
	4000	99/01	0.98 (0.97-0.99)	0.50
		95/05	0.94 (0.92-0.96)	0.54
		90/10	0.88 (0.85-0.9)	0.60
		80/20	0.84 (0.82-0.87)	0.73
		70/30	0.81 (0.78-0.84)	0.79
		60/40	0.81 (0.78-0.83)	0.79
		50/50	0.79 (0.76-0.82)	0.80
	5000	99/01	0.99 (0.98-0.99)	0.50
		95/05	0.94 (0.92-0.96)	0.58
		90/10	0.91 (0.89-0.93)	0.68
		80/20	0.85 (0.82-0.87)	0.72
		70/30	0.82 (0.79-0.84)	0.79
		60/40	0.8 (0.78-0.83)	0.80
		50/50	0.83 (0.8-0.85)	0.83

Appendix 4 - Diagnosis codes within the non-HTN group

HTN class includes ICD10 code 4019 i.e., Unspecified essential hypertension

Non-HTN class does not include any ICD10 codes beginning with 401x. The diagnoses included within this class are:

Shigella boydii	Pneumonia, organism NOS	Malignant neo colon NEC
Malign neopl prostate	Chr airway obstruct NEC	TB of limb bones-unspec
Cutaneous mycobacteria	Viral encephalitis NOS	TB limb bones-no exam
Strep sore throat	Postinflam pulm fibrosis	TB of bone NEC-unspec
Septicemia NOS	Erythema infectiosum	Malig neo pancreas NEC
Pneumococcus infect NOS	Hydronephrosis	TB of ureter-exam unkn
Subarachnoid hemorrhage	Trachoma NOS	TB of ureter-micro dx
Intracerebral hemorrhage	Early syph latent relaps	TB of ureter-histo dx
Subac scleros panenceph	TB lung infiltr-micro dx	Mal neo bronch/lung NEC
Bronchopneumonia org NOS	Malig neo tongue NOS	Malign neopl breast NEC
Sec malig neo lg bowel	TB of knee-unspec	Malig neo corpus uteri
Second malig neo liver	Benign neo skin leg	Malign neopl ovary
Sec malig neo urin NEC	Intramural leiomyoma	Malig neo bladder NEC
Sec mal neo brain/spine	Unc behav neo GI NEC	Mal neo parietal lobe
Secondary malig neo bone	Polycythemia vera	Mal neo cereb ventricle
Malignant neoplasm NOS	Hypothyroidism NOS	Malig neo brain NOS

Benign neoplasm lg bowel	Pancreatic disorder NEC	Mal neo lymph-head/neck
Ben neo liver/bile ducts	Neurohypophysis dis NEC	Mal neo lymph intra-abd
Benign neoplasm heart	Adrenal disorder NOS	Secondary malig neo lung
Benign neo skin eyelid	Testicular hypofunc NEC	Sec mal neo mediastinum
Pure hypercholesterolem	Protein-cal malnutr NOS	Second malig neo pleura
Pure hyperglyceridemia	Anemia NOS	Sec malig neo sm bowel
Hyperlipidemia NEC/NOS	Thrombocytopenia NOS	Neuropathy in diabetes
Lipoid metabol dis NOS	Wbc disease NEC	Glaucoma NOS
Gout NOS	Delirium d/t other cond	E coli septicemia
Hyperosmolality	Transient mental dis NOS	Hearing loss NOS
Hyposmolality	Mental disor NEC oth dis	Mitral insuf/aort stenos
Acidosis	Bipolor I current NOS	Mitral/aortic val insuff
Alkalosis	Obsessive-compulsive dis	Mitr/aortic mult involv
Hyperpotassemia	Dysthymic disorder	Tricuspid valve disease
Hypopotassemia	Nonpsychotic disord NOS	Mth sus Stph aur els/NOS
Chr blood loss anemia	Tobacco use disorder	Angina pectoris NEC/NOS
Iron defic anemia NOS	Bacterial meningitis NOS	Chr ischemic hrt dis NEC
B12 defic anemia NEC	Obstructiv hydrocephalus	Chr pulmon heart dis NEC
Ac posthemorrhag anemia	Paralysis agitans	Periapical abscess
Helicobacter pylori	Grand mal status	Sialoadenitis

Pericardial disease NOS	Compression of brain	Achalasia & cardiospasm
Mitral valve disorder	Trigeminal neuralgia	Esophageal stricture
Aortic valve disorder	Rupt abd aortic aneurysm	Mallory-weiss syndrome
Prim cardiomyopathy NEC	Abdom aortic aneurysm	Acq pyloric stenosis
Atriovent block complete	Thracabd anurysm wo rupt	Hernia, site NEC w obstr
Parox atrial tachycardia	Periph vascular dis NOS	Umbilical hernia
Parox ventric tachycard	Orthostatic hypotension	Diaphragmatic hernia
Cardiac arrest	Hypotension NOS	Reg enterit sm/lg intest
CHF NOS	Acute uri NOS	Ulceratve colitis unspcf
Cardiomegaly	Acute bronchitis	Allrgic gastro & colitis
Heart disease NOS	Chronic sinusitis NOS	Noninf gastroenterit NEC
Subdural hemorrhage	Vocal cord/larynx polyp	Rubella encephalitis
Intracranial hemorr NOS	Emphysema NEC	Peritoneal adhesions
Trans cereb ischemia NOS	Food/vomit pneumonitis	Rectal prolapse
Nonrupt cerebral aneurym	Pleural effusion NOS	Alcohol cirrhosis liver
Cerebrovasc disease NEC	Iatrogenic pneumothorax	Cirrhosis of liver NOS
Ruptur thoracic aneurysm	Abscess of lung	Chronic liver dis NEC
Thoracic aortic aneurysm	Pulmonary collapse	Hepatitis NOS
Blood in stool	Acute lung edema NOS	Cholangitis
Gastrointest hemorr NOS	Cervicalgia	Dis of biliary tract NEC

Human herpesvr encph NEC	Sciatica	Chronic pancreatitis
Ac kidny fail, tubr necr	Backache NOS	Pancreat cyst/pseudocyst
Acute kidney failure NOS	Other back symptoms	Hematemesis
End stage renal disease	Myalgia and myositis NOS	Patent ductus arteriosus
Chronic kidney dis NOS	Brain anomaly NEC	Intestinal anomaly NEC
Yatapoxvirus infectn NOS	Accessory auricle	Bladder exstrophy
Tanapox	Tetralogy of fallot	Down's syndrome
Stricture of ureter	Ventricular sept defect	Gonadal dysgenesis
Renal & ureteral dis NOS	Secundum atrial sept def	Hamartoses NEC
Urin tract infection NOS	Septal closure anom NEC	Congenital anomaly NOS
Noninfl dis ova/adnx NEC	Cong tricusp atres/sten	Abn plac NEC/NOS aff NB
Excessive menstruation	Cong aorta valv insuffic	Oth umbil cord compress
Cellulitis of foot	NB integument cond NEC	Exceptionally large baby
Pilonidal cyst w/o absc	Syncope and collapse	Heavy-for-date infan NEC
Diaper or napkin rash	Headache	Fetal distrs dur lab/del
Lupus erythematosus	Aphasia	NB transitory tachypnea
Other psoriasis	Epistaxis	NB cutaneous hemorrhage
Hpt C acute wo hpat coma	Tachycardia NOS	NB hemolyt dis-abo isoim
Chrnc hpt C wo hpat coma	Cardiac murmurs NEC	Neonat jaund preterm del
Hpt C w/o hepat coma NOS	Resp sys/chest symp NEC	Fetal/neonatal jaund NOS

Skin disorders NEC	Oliguria & anuria	Infant diabet mother syn
Sicca syndrome	Abn blood chemistry NEC	Neonatal dehydration
Rheumatoid arthritis	Abn find-stool contents	Neonatal hypoglycemia
Cerv spondyl w myelopath	Debility NOS	Perinatal intest perfor
Acute syphil meningitis	Fx malar/maxillary-close	NB hypothermia NEC
Contusion of chest wall	Fx malar/maxillary-open	Congenital hydrocele
Foreign body esophagus	Fx orbital floor-closed	Ch myl leuk wo achv rmsn
Injury femoral nerve	Fx orbital floor-open	Hemangioma skin
Pois-arom analgesics NEC	Fx facial bone NEC-close	Hemangioma NEC
Pois-anticonvul NEC/NOS	Fx lumbar vertebra-close	Myelodysplastic synd NOS
Pois-benzodiazepine tran	Fracture of sternum-clos	Tox dif goiter no crisis
Toxic eff ethyl alcohol	Fracture acetabulum-clos	DMII wo cmp nt st uncntr
Oth VD chlm trch unsp st	Traum pneumothorax-close	DMI wo cmp nt st uncntrl
Surg compl-heart	Traum pneumothorax-open	DMII wo cmp uncntrld
Accidental op laceration	Lac eyelid inv lacrm pas	DMII keto nt st uncntrld
Vasc comp med care NEC	Open wound of scalp	DMI keto nt st uncntrld
Second malig neo genital	Open wound of chest	DMII ketoacd uncontrold
Hdgk dis unsp xtrndl org	Open wound hand w tendon	DMI ketoacd uncontrold
Mycs fng unsp xtrndl org	Amputation finger	DMII renl nt st uncntrld
Mult mye w/o achv rmson	Abrasion head	DMII neuro nt st uncntrl

Cardiac dysrhythmias NEC	Abrasion forearm	DMII oth nt st uncntrld
Diastolc hrt failure NOS	Ulcer of heel & midfoot	DMI oth nt st uncntrld
Chr diastolic hrt fail	Osteoarthros NOS-unspec	Acute gouty arthropathy
Ill-defined hrt dis NEC	Joint symptom NEC-pelvis	Dehydration
Ocl crtd art wo infrct	Necrotizing fasciitis	Obesity NOS
Ocl mlt bi art wo infrct	Rhabdomyolysis	Morbid obesity
Crbl emblsm w infrct	Osteoporosis NOS	Anemia in neoplastic dis
Crbl art ocl NOS w infrc	Malunion of fracture	Alcohol withdrawal
Late eff CV dis-aphasia	Bone & cartilage dis NOS	Dementia w/o behav dist
Late ef-spch/lang df NEC	Forearm deformity NOS	Paranoid schizo-unspec
Late ef-hemplga side NOS	Kyphosis NOS	Schizoaffective dis NOS
Late effect CV dis NEC	Thoracogenic scoliosis	Schizoafftv dis-chr/exac
Dsct of thoracic aorta	Spin bif w hydrceph-cerv	Schizophrenia NOS-unspec
Upper extremity embolism	Spec lacrimal pass anom	Bipol I currnt manic NOS
Bleed esoph var oth dis	Ex ear anm NEC-impr hear	Anxiety state NOS
Iatrogenc hypotnsion NEC	Ostium primum defect	Conversion disorder
Obs chr bronc w(ac) exac	Cong pulmon valve stenos	Borderline personality
Ext asthma w status asth	Cong heart anomaly NEC	Ac alcohol intox-contin
Chronic obst asthma NOS	Great vein anomaly NEC	Alcoh dep NEC/NOS-unspec
Asthma NOS	Cerebrovascular anomaly	Alcoh dep NEC/NOS-contin

Asthma NOS w (ac) exac	Spinal vessel anomaly	Opioid dependence-unspec
Acute respiratry failure	Biliary & liver anom NEC	Opioid dependence-contin
Other pulmonary insuff	Hypospadias	Drug depend NOS-unspec
Acute & chronc resp fail	Obst def ren plv&urt NEC	Alcohol abuse-unspec
Tracheostomy - mech comp	Congn anoml abd wall NEC	Alcohol abuse-continuous
Other esophagitis	Cong skin pigment anomal	Cocaine abuse-unspec
Ulc esophagus w/o bleed	Nox sub NEC aff NB/fetus	Amphetamine abuse-unspec
Esophageal reflux	Lt-for-dates 1750-1999g	Drug abuse NEC-unspec
Barrett's esophagus	Lt-for-dates 2000-2499g	Attn deficit w hyperact
Chr stomach ulc w hem	Preterm NEC 1750-1999g	Obstructive sleep apnea
Stomach ulcer NOS	Preterm NEC 2000-2499g	Dementia w Lewy bodies
Chr duoden ulcer w hem	Preterm NEC 2500+g	Amyotrophic sclerosis
Duodenal ulcer NOS	33-34 comp wks gestation	Psymotr epil w/o int epi
Chr marginal ulc w perf	35-36 comp wks gestation	Part epil w/o intr epil
Oth spf gastrt w hmrhg	37+ comp wks gestation	Epilep NOS w/o intr epil
Gstr/ddnts NOS w/o hmrhg	Injuries to scalp NEC	Othr migrne wo ntrc mgrn
Gstr/ddnts NOS w hmrhg	Primary apnea of newborn	Migrne unsp wo ntrc mgrn
Duodenitis w/o hmrhg	Cyanotic attack, newborn	Rheumatic heart failure
Gastroduodenal dis NEC	Resp failure of newborn	Hy kid NOS w cr kid I-IV
Intestinal obstruct NEC	Resp prob after brth NEC	Hyp kid NOS w cr kid V

Dvrtcli colon w/o hmrhg	Bacteremia of newborn	AMI anterior wall, init
Dvrtclo colon w hmrhg	Neonatal tachycardia	AMI inferolateral, init
Constipation NOS	Meconium staining	AMI inferopost, initial
Peritonitis (acute) gen	Perinatal condition NEC	AMI inferior wall, init
Peritonitis NEC	Other alter consciousnes	True post infarct, init
Ulceration of intestine	Convulsions NEC	Subendo infarct, initial
Perforation of intestine	Sleep apnea NOS	Subendo infarct, subseq
Angio intes w hmrhg	Cardiogenic shock	AMI NEC, initial
Cholelith w ac cholecyst	Septic shock	AMI NOS, initial
Cholelithiasis NOS	Shock w/o trauma NEC	AMI NOS, subsequent
Gall&bil cal w/oth w obs	Chest pain NOS	Cor ath unsp vsl ntv/gft
Nephritis NOS in oth dis	Nausea with vomiting	Crnry athrscl natve vssl
Ac pyelonephritis NOS	Diarrhea	Crn ath atlg vn bps grft
Neurogenic bladder NOS	Retention urine NOS	Cor ath artry bypas grft
BPH w/o urinary obs/LUTS	Urinary frequency	Aneurysm coronary vessel
Legal abort w hemorr-inc	Drop, hematocrit, precip	Atrioven block-mobitz ii
Mild/NOS preeclamp-p/p	Abnrml coagultion prfile	Conduction disorder NEC
Anemia-delivered w p/p	Hypoxemia	Atrial fibrillation
CV dis NEC-antepartum	Cl skl vlt fx/cerebr lac	Atrial flutter
Postpa hem NEC-del w p/p	Cl skl base fx/cereb lac	Ventricular fibrillation

P/p coag def-del w p/p	Cl skl base fx/menin hem	Sinoatrial node dysfunct
Peripartum card-postpart	Cl skul base fx w/o coma	Abn react-anastom/graft
Brain lacer NEC w/o coma	Cl skl fx NEC/mening hem	Abn reac-organ rem NEC
Subarach hem-brief coma	Cl skul fx NEC-deep coma	Abn react-surg proc NEC
Subarach hem-deep coma	Cl skl w oth fx-coma NOS	Abn react-cardiac cath
Subarach hem-coma NOS	Fx c2 vertebra-closed	Abn react-radiotherapy
Traumatic subdural hem	Fx mult cervical vert-cl	Abn react-procedure NOS
Subdural hem w/o coma	C5-c7 fx-cl/ant cord syn	Fall on stair/step NEC
Traumatic brain hem NEC	Fracture three ribs-clos	Fall from ladder
Brain hem NEC-coma NOS	Fracture seven ribs-clos	Diving accident
Heart contusion-closed	Fx tibia NOS-closed	Fall-1 level to oth NEC
Lung contusion-closed	Fx tibia w fibula NOS-cl	Fall from slipping NEC
Duodenum injury-closed	Disloc 2nd cerv vert-cl	Fall NOS
Sigmoid colon inj-closed	Sprain of ankle NOS	Resp obstr-food inhal
Liver hematoma/contusion	Brain laceration NEC	FB entering oth orifice
Liver lacerat unspcf cls	Comp-oth vasc dev/graft	Struck by falling object
Liver injury NEC-closed	Hemorrhage complic proc	Woodworking machine acc
Spleen injury NOS-closed	Hematoma complic proc	Machinery accident NEC
Spleen hematoma-closed	Other postop infection	Acc-cutting instrum NEC
Spleen disruption-clos	Non-healing surgcl wound	Hunting rifle accident

Spleen injury NEC-closed	Oth spcf cmplc procd NEC	Adv eff cephalosporin
Open wound of forehead	Mv collision NOS-driver	Adv eff antineoplastic
Open wound of jaw	Mv collision NOS-pasngr	Adv eff opiates
Open wound of face NEC	Mv-oth veh coll-driver	Adv eff analgesic NOS
Poisoning-opiates NEC	Mv coll w ped-ped cycl	Adv eff sedat/hypnot NEC
Pois-propionic acid derv	Mv coll w pedest-pedest	Adv eff coronary vasodil
Severe sepsis	Loss control mv acc-driv	Poison-analgesics
SIRS-noninf w/o ac or ds	Loss control mv acc-psgr	Poison-drug/medicin NEC
Malfunc prosth hrt valve	Traffic acc NOS-driver	Poison-solid/liquid NEC
Periprosthetc osteolysis	Ped cycl acc-ped cyclist	Unarmed fight or brawl
React-int pros devic NEC	Accid in recreation area	Assault-cutting instr
Comp-heart valve prosth	Acc poisn-benzdiaz tranq	Assault-striking w obj
Comp-oth cardiac device	Abn react-artif implant	Assault NOS
Undeter pois-sol/liq NEC	Prsnl hst colonic polyps	Undeterm pois-analgesics
Need prphyl vc vrl hepat	Prsnl hst ot spf dgst ds	Undeterm pois-psychotrop
Asymp hiv infectn status	Trnspl status-pancreas	Heart valve replac NEC
Hx of colonic malignancy	History of tobacco use	Joint replaced hip
Hx-bronchogenic malignan	Family hx-breast malig	Joint replaced knee
Hx of breast malignancy	Fam hx-ischem heart dis	Status cardiac pacemaker
Hx-uterus malignancy NEC	Fam hx-diabetes mellitus	Acq absnce breast/nipple

Hx-prostatic malignancy	NB obsrv suspct infect	Acquired absence kidney
Hx of bladder malignancy	NB obs genetc/metabl cnd	Acq absence of lung
Hx-lymphatic malign NEC	NB obsrv oth suspct cond	Aortocoronary bypass
Hx-malig skin melanoma	Single lb in-hosp w/o cs	Status-post ptca
Hx-skin malignancy NEC	Single lb in-hosp w cs	Routine circumcision
Hx of brain malignancy	Singl livebrn-before adm	Long-term use anticoagul
Hx of affective disorder	Twin-mate lb-hosp w/o cs	Long-term use of insulin
Personal histry malaria	Twin-mate lb-in hos w cs	Wait adm to oth facility
Hx-ven thrombosis/embols	Twin-mate sb-hosp w cs	No proc/contraindication
Hx-circulatory dis NEC	Oth mult lb-in hosp w cs	No proc/patient decision
Observ-accident NEC	Kidney transplant status	Exam-clincal trial

Appendix 5 - Annotator Disagreements

Class	Disagreements	Final Decisions
Relevant (options are Relevant, Not Relevant, Negated)	painful memories, pain from talking about the past, finding things painful to accept	Not relevant - we are not including instances of mental pain, only physical pain.
	Taking pain relief meds, pain killers	Relevant, as it indicates they are in pain
	Referred for pain management	Relevant, as it indicates they are in pain
	Mention of 'burns' as a noun rather than describing pain. Ex - he has severe burns (rather than something like burning pain)	Not relevant, as it is not about pain
	Fear of pain	Not relevant, as it is hypothetical
	Inflicting pain on others (such as by hitting them)	Not relevant
Pain Quality	'various pains' as a character	NA

(options are Chronic, Other, NA)	period pain	Other
Anatomy (options are Mentioned, NA)	pain on left side of body, pain all over the body, body pain	Mentioned
	somatic pain	NA
	period pain	NA
Pain Management (options are Medication/Other/NA)	“medication” i.e., not an actual drug name	Medication
	“Pain management team”, “under the pain management team”, “Known to the pain management team”	Other
	recommendation to visit the GP about some pain the patient is experiencing	Other
	side effect of psychotic medication is shooting pains in legs	NA, as the medication is not for pain management

Further details on the annotation and adjudication guidelines can be found at:

https://docs.google.com/document/d/1-zpPhk26i62a22miSNTNxRr6XtAkQyXYDiAe98dZkXU/edit?usp=sharing

Appendix 6 - Environmental Impact

Recent advances in technology and the rise of data-intensive research has led to the development of powerful computational tools, such as GPUs, which have significantly accelerated the progress within the field of NLP. These tools have helped produce powerful language models. However, it is essential to acknowledge that while these computational tools offer unprecedented capabilities, they also exert substantial demands on energy resources, leading to carbon emissions and subsequent environmental consequences.

GPUs were used in the development of the classifier models presented in this project, especially for the transformer-based BERT models. It is essential to recognise the impact of the use of such resources on climate change and the carbon footprint in general. Carbon emissions associated with the utilisation of GPUs for the development of transformer-based models are reported here, not only for transparency but also to promote sustainability in research practices. The intention of quantifying the carbon emissions is to provide an overview of the environmental implications of such research projects.

An MIT technology review highlighted that the computational and environmental costs of training an AI model grow proportionally to the model's size, compounded further by the tuning steps added to improve the performance scores of the models. The review also claimed that training BERT models (with 110M parameters) have a carbon footprint of about 1,400 pounds of carbon dioxide, which is equivalent to a round-trip trans-American flight (Hao, 2019).

Since this project did not undertake training any BERT models from scratch, and since the implications of fine-tuning a model are significantly lower, the impact on carbon emissions from this project is not noteworthy. A tool[31], developed by Lacoste et al. (2019), was used to

[31] https://mlco2.github.io/impact/#publish

calculate the carbon emissions from the models developed in this project (Lacoste et al., 2019).

Experiments in this project were conducted within NCasT4_v3-series Azure virtual machine infrastructure in the region “UK South”, which has a carbon efficiency of 0.62 kgCO_2eq/kWh. A cumulative of approximately 100 hours of computation was performed on the hardware of type Nvidia Tesla T4. Total emissions are estimated to be 4.34 kgCO_2eq, of which 100 per cent were directly offset by the cloud provider, Azure. 4.34 kgCO_2eq is equivalent to 17.5km driven by an average ICE (internal combustion engine) car, or 2.17 kg of coal burned.

Appendix 7 – The RECORD statement

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	(a) Title (b) Abstract	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Abstract
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Introduction - Background Rationale		
Objectives	3	State specific objectives, including any prespecified hypotheses	Introduction - Objectives		
Methods					
Study Design	4	Present key elements of study design early in the paper	Methods		
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods - Setting and Variables		
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case	Methods - Participants	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Methods - Variables	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods - Variables		

Bias	9	Describe any efforts to address potential sources of bias	Methods - Variables		
Study size	10	Explain how the study size was arrived at	Methods - Participants		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Methods - Variables		
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Methods - Setting and Descriptive Statistics		
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.	Methods - Setting - Ethical Approval
				RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	Methods - Variables
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Methods - Setting
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	Results	RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Methods
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)	Results		
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure	Results		

		category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Results		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Results		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Discussion		
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Discussion and Conclusion
Interpretation	20	Give a cautious overall interpretation of results considering objectives,			
		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion		
Generalisability	21	Discuss the generalisability (external validity) of the study results	Discussion		
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Acknolwledgements and disclosure of interests		
Accessibility of protocol, raw data, and programming code		..	Data Availability Statement	RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Data Availability Statement

*Reference: Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; in press.

*Checklist is protected under Creative Commons Attribution (CC BY) license.

12 References

Abhyankar, S., Demner-Fushman, D., Callaghan, F. M., & McDonald, C. J. (2014). Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association*, *21*(5), 801–807. https://doi.org/10.1136/amiajnl-2013-001915

Abplanalp, S. J., Mueser, K. T., & Fulford, D. (2020). The role of physical pain in global functioning of people with serious mental illness. *Schizophrenia Research*, *222*, 423–428. https://doi.org/10.1016/j.schres.2020.03.062

Advance HE. (2003, August 1). *Quality in Qualitative Evaluation: A framework for assessing research evidence* . https://www.advance-he.ac.uk/knowledge-hub/quality-qualitative-evaluation-framework-assessing-research-evidence

Agarwal, A. (2015). *Reddit Word Embeddings*. https://kaggle.com/alaap29/reddit-word-embeddings

Agarwal, K., Eftimov, T., Addanki, R., Choudhury, S., Tamang, S., & Rallo, R. (2019). Snomed2Vec: Random Walk and Poincar\'e Embeddings of a Clinical Knowledge Base for Healthcare Analytics. *ArXiv:1907.08650 [Cs, Stat]*.

Aguggia, M. (2003). Neurophysiology of pain. *Neurological Sciences*, *24 Suppl 2*, S57-60. https://doi.org/10.1007/s100720300042

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. https://doi.org/10.18653/v1/W19-1909

Alshahrani, M., & Hoehndorf, R. (2018). Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, *34*(17), i901–i907. https://doi.org/10.1093/bioinformatics/bty559

Alshahrani, M., Khan, M. A., Maddouri, O., Kinjo, A. R., Queralt-Rosinach, N., & Hoehndorf, R. (2017). Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics*, *33*(17), 2723–2730. https://doi.org/10.1093/bioinformatics/btx275

Alshahrani, M., Thafar, M. A., & Essack, M. (2021). Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ. Computer Science*, *7*, e341. https://doi.org/10.7717/peerj-cs.341

American Psychiatric Association. (2020). *Chronic Pain and Mental Health Often Interconnected*. https://www.psychiatry.org/news-room/apa-blogs/chronic-pain-and-mental-health-interconnected

AmpliGraph. (2019a). *Performance — AmpliGraph 2.0.0 documentation*. https://docs.ampligraph.org/en/2.0.0/experiments.html

AmpliGraph. (2019b, March). *AmpliGraph — AmpliGraph 2.0.0 documentation*. https://docs.ampligraph.org/en/1.1.0/

Aro, S., Aro, H., & Keskimäki, I. (1995). Socio-economic mobility among patients with schizophrenia or major affective disorder. A 17-year retrospective follow-up. *The British Journal of Psychiatry*, *166*(6), 759–767. https://doi.org/10.1192/bjp.166.6.759

Azzam, S., Humphreys, K., Gaizauskas, R., & Wilks, Y. (1999). Using a language independent domain model for multilingual information extraction. *Applied Artificial Intelligence*, *13*(7), 705–724. https://doi.org/10.1080/088395199117252

Bacco, L., Russo, F., Ambrosio, L., D'Antoni, F., Vollero, L., Vadalà, G., Dell'Orletta, F., Merone, M., Papalia, R., & Denaro, V. (2022). Natural language processing in low back pain and spine diseases: A systematic review. *Frontiers in Surgery*, *9*, 957085. https://doi.org/10.3389/fsurg.2022.957085

Bair, M. J., Robinson, R. L., Katon, W., & Kroenke, K. (2003). Depression and pain comorbidity: a literature review. *Archives of Internal Medicine*, *163*(20), 2433–2445. https://doi.org/10.1001/archinte.163.20.2433

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., & Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, *380*(9836), 37–43. https://doi.org/10.1016/S0140-6736(12)60240-2

Baud, R. H., Rassinoux, A. M., & Scherrer, J. R. (1992). Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, *31*(2), 117–125. https://doi.org/10.1055/s-0038-1634865

Baughman, K. R., Bonfine, N., Dugan, S. E., Adams, R., Gallagher, M., Olds, R. S., Piatt, E., & Ritter, C. (2016). Disease Burden Among Individuals with Severe Mental Illness in a Community Setting. *Community Mental Health Journal*, *52*(4), 424–432. https://doi.org/10.1007/s10597-015-9973-2

Bear Don't Walk Iv, O. J., Sun, T., Perotte, A., & Elhadad, N. (2021). Clinically relevant pretraining is all you need. *Journal of the American Medical Informatics Association*, *28*(9), 1970–1976. https://doi.org/10.1093/jamia/ocab086

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Hemkens, L. G., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M., & RECORD Working Committee. (2016). [The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement]. *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen*, *115–116*, 33–48. https://doi.org/10.1016/j.zefq.2016.07.010

Bendayan, R., Kraljevic, Z., Shaari, S., Das-Munshi, J., Leipold, L., Chaturvedi, J., Mirza, L., Aldelemi, S., Searle, T., Chance, N., Mascio, A., Skiada, N., Wang, T., Roberts, A., Stewart, R., Bean, D., & Dobson, R. (2022). Mapping multimorbidity in individuals with schizophrenia and bipolar disorders: evidence from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register. *BMJ Open*, *12*(1), e054414. https://doi.org/10.1136/bmjopen-2021-054414

Bendayan, R., Wu, H., Kraljevic, Z., Stewart, R., Searle, T., Chaturvedi, J., Das-Munshi, J., Ibrahim, Z., Mascio, A., Roberts, A., Bean, D., & Dobson, R. (2020). Identifying physical health comorbidities in a cohort of individuals with severe mental illness: An application of SemEHR. *ArXiv*. https://doi.org/10.48550/arxiv.2002.08901

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bennett, C. (2020). *What you need to know about pain and mental health in Australia*. Pain Australia. https://www.painaustralia.org.au/media-document/blog-1/blog-2020/blog-2019/what-you-need-to-know-about-pain-and-mental-health-in-australia

Benson, T. (2010). SNOMED CT. In *Principles of health interoperability HL7 and SNOMED* (pp. 189–215). Springer London. https://doi.org/10.1007/978-1-84882-803-2_12

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102. https://doi.org/10.18653/v1/W17-1612

Birman-Deych, E., Waterman, AD., Yan, Y., Nilasena, DS., Radford, MJ., & Gage, BF. (2005). Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors. Medical Care, 43(5), 480–485. http://www.jstor.org/stable/3768402

Bianchi, F., Rossiello, G., Costabello, L., Palmonari, M., & Minervini, P. (2020). *Knowledge Graph Embeddings and Explainable AI*.

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards Large-scale Twitter Mining for Drug-related Adverse Events. *SHB'12 : Proceedings of the 2012 ACM International Workshop on Smart Health and Wellbeing : October 29, 2012, Maui, Hawaii, USA. International Workshop on Smart Health and Wellbeing (2012 : Maui, Hawaii)*, *2012*, 25–32. https://doi.org/10.1145/2389707.2389713

Biondo, F., Jewell, A., Pritchard, M., Aarsland, D., Steves, C. J., Mueller, C., & Cole, J. H. (2022). Brain-age is associated with progression to dementia in memory clinic patients. *NeuroImage. Clinical*, *36*, 103175. https://doi.org/10.1016/j.nicl.2022.103175

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Birgenheir, D. G., Ilgen, M. A., Bohnert, A. S. B., Abraham, K. M., Bowersox, N. W., Austin, K., & Kilbourne, A. M. (2013). Pain conditions among veterans with schizophrenia or bipolar disorder. *General Hospital Psychiatry*, *35*(5), 480–484. https://doi.org/10.1016/j.genhosppsych.2013.03.019

Biswas, S., Mitra, P., & Rao, K. S. (2021). Relation prediction of co-morbid diseases using knowledge graph completion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 708–717. https://doi.org/10.1109/TCBB.2019.2927310

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, 106. https://doi.org/10.1186/1471-2105-14-106

Bleuler, E. (1988). *Textbook of psychiatry* (Special ed.). New York : Classics of Psychiatry & Behavioral Sciences Library, 1988.

Blumer, D., & Heilbronn, M. (1982). Chronic pain as a variant of depressive disease: the pain-prone disorder. *The Journal of Nervous and Mental Disease*, *170*(7), 381–406. https://doi.org/10.1097/00005053-198207000-00001

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267-70. https://doi.org/10.1093/nar/gkh061

Boe, B. (2012). *PRAW: The Python Reddit API Wrapper.* https://github.com/praw-dev/praw/

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bonnot, O., Anderson, G. M., Cohen, D., Willer, J. C., & Tordjman, S. (2009). Are patients with schizophrenia insensitive to pain? A reconsideration of the question. *The Clinical Journal of Pain*, *25*(3), 244–252. https://doi.org/10.1097/AJP.0b013e318192be97

Boot, A. B., Tjong Kim Sang, E., Dijkstra, K., & Zwaan, R. A. (2019). How character limit affects language usage in tweets. *Palgrave Communications*, *5*(1), 76. https://doi.org/10.1057/s41599-019-0280-3

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). *Translating Embeddings for Modeling Multi-relational Data*. 9.

Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., Velupillai, S., & Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. *BMJ Open*, *12*(2), e052911. https://doi.org/10.1136/bmjopen-2021-052911

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). *#LancsBox [software]* (v.5.x) [Computer software].

Bridges, S. (2014). Chapter 2: Mental health problem. In *Health Survey* (Vol. 1).

Brooks, J. M., Umucu, E., Huck, G. E., Fortuna, K., Sánchez, J., Chiu, C., & Bartels, S. J. (2018). Sociodemographic characteristics, health conditions, and functional impairment

among older adults with serious mental illness reporting moderate-to-severe pain. *Psychiatric Rehabilitation Journal*, *41*(3), 224–233. https://doi.org/10.1037/prj0000316

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*.

Bui, D. D. A., & Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, *21*(5), 850–857. https://doi.org/10.1136/amiajnl-2013-002411

Cai, J., & Zeng, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics*, *60*(4), 1015–1024. https://doi.org/10.1111/j.0006-341X.2004.00257.x

Campbell, C. M., & Edwards, R. R. (2012). Ethnic differences in pain and pain management. *Pain Management*, *2*(3), 219–230. https://doi.org/10.2217/pmt.12.7

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). With Little Power Comes Great Responsibility. *ArXiv*. https://doi.org/10.48550/arxiv.2010.06595

Carey, D. J., Fetterolf, S. N., Davis, F. D., Faucett, W. A., Kirchner, H. L., Mirshahi, U., Murray, M. F., Smelser, D. T., Gerhard, G. S., & Ledbetter, D. H. (2016). The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genetics in Medicine*, *18*(9), 906–913. https://doi.org/10.1038/gim.2015.187

Carlson, L. A., & Hooten, W. M. (2020). Pain-Linguistics and Natural Language Processing. *Mayo Clinic Proceedings. Innovations, Quality & Outcomes*, *4*(3), 346–347. https://doi.org/10.1016/j.mayocpiqo.2020.01.005

Carver, A. C., & Foley, K. M. (2003). *Types of Pain*.

Catalao, R., Ashworth, M., Stewart, R., Hatch, S. L., & Howard, L. M. (2021). Racial and ethnic disparities in multimorbidity in non-pregnant women of reproductive age in Lambeth, UK: a data linkage study. *The Lancet*, *398*, S31. https://doi.org/10.1016/S0140-6736(21)02574-5

Chang, D., Balažević, I., Allen, C., Chawla, D., Brandt, C., & Taylor, R. A. (2020). Benchmark and best practices for biomedical knowledge graph embeddings. *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, *2020*, 167–176. https://doi.org/10.18653/v1/2020.bionlp-1.18

Chang, D., Lin, E., Brandt, C., & Taylor, R. A. (2021). Incorporating domain knowledge into language models by using graph convolutional networks for assessing semantic textual similarity: Model development and performance comparison. *JMIR Medical Informatics*, *9*(11), e23101. https://doi.org/10.2196/23101

Chang, E., Rashid, M. H., Lin, P.-J., Zhao, C., Demberg, V., Shi, Y., and Chandra, V. (2023). Revisiting Sample Size Determination in Natural Language Understanding. ArXiv. https://doi.org/10.48550/arxiv.2307.00374

Chaturvedi, J., Mascio, A., Velupillai, S. U., & Roberts, A. (2021). Development of a lexicon for pain. *Frontiers in Digital Health*, *3*, 778305. https://doi.org/10.3389/fdgth.2021.778305

Chaturvedi, J., Velupillai, S., Stewart, R., and Roberts, A. (2024). Identifying mentions of pain in mental health records text: A natural language processing approach. Studies in Health Technology and Informatics, 310, 695–699. https://doi.org/10.3233/SHTI231054

Chaturvedi, J., Viani, N., Velupillai, S., & Roberts, A. (2019). Analysis and Annotation of temporal information related to medications in EHRs. *Unpublished*. https://doi.org/10.13140/rg.2.2.25224.67846

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10.

Chou, W.-Y. S., Prestin, A., & Kunath, S. (2014). Obesity in social media: a mixed methods analysis. *Translational Behavioral Medicine*, *4*(3), 314–323. https://doi.org/10.1007/s13142-014-0256-1

Colling, C., Mueller, C., Perera, G., Funnell, N., Sauer, J., Harwood, D., Stewart, R., & Bishara, D. (2020). “Real time” monitoring of antipsychotic prescribing in patients with dementia: a study using the Clinical Record Interactive Search (CRIS) platform to enhance safer prescribing. *BMJ Open Quality*, *9*(1). https://doi.org/10.1136/bmjoq-2019-000778

Coppersmith, G., Leary, R., Wood, T., & Whyne, E. (2015). Quantifying Suicidal Ideation via Language Usage on Social Media. *Joint Statistics Meetings Proceedings, Statistical Computing Section*, 15.

Costabello, L., Pai, S., Le van, C., McGrath, R., McCarthy, N., & Tabacof, P. (2019). *AmpliGraph: a Library for Representation Learning on Knowledge Graphs* [Computer software].

Cruse, A. (2004). *Meaning in language: An introduction to semantics and pragmatics.*

Cuomo, A., Bimonte, S., Forte, C. A., Botti, G., & Cascella, M. (2019). Multimodal approaches and tailored therapies for pain management: the trolley analgesic model. *Journal of Pain Research*, *12*, 711–714. https://doi.org/10.2147/JPR.S178910

Dagenais, S., Caro, J., and Haldeman, S. (2008). A systematic review of low back pain cost of illness studies in the United States and internationally. The Spine Journal, 8(1), 8–20. https://doi.org/10.1016/j.spinee.2007.10.005

Dave, A. D., Ruano, G., Kost, J., and Wang, X. (2022). Automated Extraction of Pain Symptoms: A Natural Language Approach using Electronic Health Records. Pain Physician, 25(2), E245–E254.

Davis, K. A. S., Mueller, C., Ashworth, M., Broadbent, M., Jewel, A., Molokhia, M., Perera, G., & Stewart, R. J. (2021). What gets recorded, counts: dementia recording in primary care compared with a specialist database. *Age and Ageing*, *50*(6), 2206–2213. https://doi.org/10.1093/ageing/afab164

DE Hert, M., Correll, C. U., Bobes, J., Cetkovich-Bakmas, M., Cohen, D., Asai, I., Detraux, J., Gautam, S., Möller, H.-J., Ndetei, D. M., Newcomer, J. W., Uwakwe, R., & Leucht, S. (2011). Physical illness in patients with severe mental disorders. I. Prevalence, impact of medications and disparities in health care. *World Psychiatry : Official Journal of the World*

Psychiatric Association (WPA), *10*(1), 52–77. https://doi.org/10.1002/j.2051-5545.2011.tb00014.x

DELİCE, A. (2010). The Sampling Issues in Quantitative Research. *Educational Sciences: Theory & Practice*, *10*(4), 18.

Denaxas, S. C., & Morley, K. I. (2015). Big biomedical data and cardiovascular disease research: opportunities and challenges. *European Heart Journal - Quality of Care and Clinical Outcomes*, *1*(1), 9–16. https://doi.org/10.1093/ehjqcco/qcv005

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/10.48550/arxiv.1810.04805

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2098–2110. https://doi.org/10.1145/2858036.2858207

Dixon, J., Sanderson, C., Elliott, P., Walls, P., Jones, J., & Petticrew, M. (1998). Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. *Journal of Public Health Medicine*, *20*(1), 63–69. https://doi.org/10.1093/oxfordjournals.pubmed.a024721

Dorflinger, L. M., Gilliam, W. P., Lee, A. W., & Kerns, R. D. (2014). Development and application of an electronic health record information extraction tool to assess quality of pain management in primary care. *Translational Behavioral Medicine*, *4*(2), 184–189. https://doi.org/10.1007/s13142-014-0260-5

Dorrington, S., Carr, E., Polling, C., Stevelink, S., Ashworth, M., Roberts, E., Broadbent, M., Hatch, S., Madan, I., & Hotopf, M. (2021). Health condition at first fit note and number of fit notes: a longitudinal study of primary care records in south London. *BMJ Open*, *11*(3), e043889. https://doi.org/10.1136/bmjopen-2020-043889

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, *125*, 37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

Ehrenberg, A. S. C. (2000). Data Reduction. *Journal of Empirical Generalisations in Marketing Science*, *5*(1), 290–291.

Ehrenstein, V., Kharrazi, H., Lehmann, H., & Taylor, C. O. (2019). *Obtaining Data From Electronic Health Records*. Agency for Healthcare Research and Quality (US).

Ehrlinger, L., & Wöß, W. (2016). *Towards a Definition of Knowledge Graphs*. International Conference on Semantic Systems.

Eisendrath, S. J. (1995). Psychiatric aspects of chronic pain. *Neurology*, *45*(12 Suppl 9), S26-34; discussion S35. https://doi.org/10.1212/wnl.45.12_suppl_9.s26

Ennajari, H., Bouguila, N., & Bentahar, J. (2022). Knowledge-enhanced Spherical Representation Learning for Text Classification. In A. Banerjee, Z.-H. Zhou, E. E. Papalexakis, & M. Riondato (Eds.), *Proceedings of the 2022 SIAM international conference on data mining (SDM)* (pp. 639–647). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611977172.72

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

Fang, Y., Zhao, X., Tan, Z., & Xiao, W. (2018). Relational Knowledge Prediction via Dynamic Bi-Mode Embedding. *IEEE Access : Practical Innovations, Open Solutions*, *6*, 25715–25723. https://doi.org/10.1109/ACCESS.2018.2832165

Faris, H., Faris, M., Habib, M., & Alomari, A. (2022). Automatic symptoms identification from a massive volume of unstructured medical consultations using deep neural and BERT models. *Heliyon*, *8*(6), e09683. https://doi.org/10.1016/j.heliyon.2022.e09683

Fernandes, A. C., Cloete, D., Broadbent, M. T. M., Hayes, R. D., Chang, C.-K., Jackson, R. G., Roberts, A., Tsang, J., Soncul, M., Liebscher, J., Stewart, R., & Callard, F. (2013). Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Medical Informatics and Decision Making*, *13*, 71. https://doi.org/10.1186/1472-6947-13-71

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. *Scientific Reports*, *8*(1), 7426. https://doi.org/10.1038/s41598-018-25773-2

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*, 8. https://doi.org/10.1186/1472-6947-12-8

Fillingim, Roger B. (2017). Sex, gender, and pain. In *Principles of Gender-Specific Medicine* (pp. 481–496). Elsevier. https://doi.org/10.1016/B978-0-12-803506-1.00038-3

Fillingim, R B. (2000). Sex, gender, and pain: women and men really are different. *Current Review of Pain*, *4*(1), 24–30. https://doi.org/10.1007/s11916-000-0006-6

Fishman, S. M. (2007). Recognizing pain management as a human right: a first step. Anesthesia and Analgesia, 105(1), 8–9. https://doi.org/10.1213/01.ane.0000267526.37663.41

Fitzner, K., & Heckinger, E. (2010). Sample size calculation and power analysis: a quick review. *The Diabetes Educator*, *36*(5), 701–707. https://doi.org/10.1177/0145721710380791

Fodeh, S. J., Finch, D., Bouayad, L., Luther, S. L., Ling, H., Kerns, R. D., & Brandt, C. (2018). Classifying clinical notes with pain assessment using machine learning. *Medical & Biological Engineering & Computing*, *56*(7), 1285–1292. https://doi.org/10.1007/s11517-017-1772-1

Ford, E., Oswald, M., Hassan, L., Bozentko, K., Nenadic, G., & Cassell, J. (2020). Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. *Journal of Medical Ethics*, *46*(6), 367–377. https://doi.org/10.1136/medethics-2019-105472

Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., & Song, M. (2019). Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations. *Journal of Medical Internet Research*, *21*(6), e12876. https://doi.org/10.2196/12876

Frayne, S. M., Halanych, J. H., Miller, D. R., Wang, F., Lin, H., Pogach, L., Sharkansky, E. J., Keane, T. M., Skinner, K. M., Rosen, C. S., & Berlowitz, D. R. (2005). Disparities in diabetes care: impact of mental illness. *Archives of Internal Medicine*, *165*(22), 2631–2638. https://doi.org/10.1001/archinte.165.22.2631

Friedman, C., Rindflesch, T. C., & Corn, M. (2013). Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of Biomedical Informatics*, *46*(5), 765–773. https://doi.org/10.1016/j.jbi.2013.06.004

Fung, K. W., Bodenreider, O., Aronson, A. R., Hole, W. T., & Srinivasan, S. (2007). Combining lexical and semantic methods of inter-terminology mapping using the UMLS. *Studies in Health Technology and Informatics*, *129*(Pt 1), 605–609.

Gaizauskas, R., & Humphreys, K. (1997). Using a semantic network for information extraction. *Natural Language Engineering*, *3*(2), 147–169. https://doi.org/10.1017/S1351324997001769

Garla, V., Taylor, C., & Brandt, C. (2013). Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *Journal of Biomedical Informatics*, *46*(5), 869–875. https://doi.org/10.1016/j.jbi.2013.06.014

Ghannay, S., Favre, B., Estève, Y., & Camelin, N. (2016). Word Embedding Evaluation and Combination. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 300–305.

Goldberg, D. S., & McGee, S. J. (2011). Pain as a global public health priority. *BMC Public Health*, *11*, 770. https://doi.org/10.1186/1471-2458-11-770

GOV.UK. (2019). *English indices of deprivation 2019*. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Govind, R., de Freitas, D. F., Pritchard, M., Khondoker, M., Teo, J. T., Stewart, R., Hayes, R. D., & MacCabe, J. H. (2022). COVID-related hospitalization, intensive care treatment, and all-cause mortality in patients with psychosis and treated with clozapine. *European Neuropsychopharmacology*, *56*, 92–99. https://doi.org/10.1016/j.euroneuro.2022.01.007

Gra.fo. (2020). *Gra.fo* . A Data.World Company. https://gra.fo/

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. *ArXiv*. https://doi.org/10.48550/arxiv.2008.05756

Green, C. R., Anderson, K. O., Baker, T. A., Campbell, L. C., Decker, S., Fillingim, R. B., Kalauokalani, D. A., Lasch, K. E., Myers, C., Tait, R. C., Todd, K. H., & Vallerand, A. H. (2003). The unequal burden of pain: confronting racial and ethnic disparities in pain. *Pain Medicine*, *4*(3), 277–294. https://doi.org/10.1046/j.1526-4637.2003.03034.x

Groenewald, C. B., Essner, B. S., Wright, D., Fesinmeyer, M. D., & Palermo, T. M. (2014). The economic costs of chronic pain among a cohort of treatment-seeking adolescents in the United States. *The Journal of Pain*, *15*(9), 925–933. https://doi.org/10.1016/j.jpain.2014.06.002

Gureje, O., Von Korff, M., Simon, G. E., & Gater, R. (1998). Persistent pain and well-being: a World Health Organization Study in Primary Care. *The Journal of the American Medical Association*, *280*(2), 147–151. https://doi.org/10.1001/jama.280.2.147

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, ontology, and precision medicine. *The New England Journal of Medicine*, *379*(15), 1452–1462. https://doi.org/10.1056/NEJMra1615014

Hafezparast, N., Bragan Turner, E., Dunbar-Rees, R., Vusirikala, A., de La Morinière, V., Yeo, K., Dodhia, H., Durbaba, S., Shetty, S., & Ashworth, M. (2023). Identifying populations with chronic pain in primary care: developing an algorithm and logic rules applied to coded primary care diagnostic and medication data. *BMC Primary Care*.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Hao, K. (2019, June 6). *Training a single AI model can emit as much carbon as five cars in their lifetimes* . MIT Technology Review. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing

errors. *Statistics in Medicine*, *15*(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Harris, Z. S. (1968). Mathematical structures of language. *Wiley Interscience*.

Harris, Z. S. (1982). A grammar of English on mathematical principles. *Wiley & Sons*.

Harris, Z. S. (1991). Theory of language and information: a mathematical approach. *Clarendon Press*.

Hastie, T., Friedman, J., & Tibshirani, R. (2001a). Support Vector Machines and Flexible Discriminants. In T. Hastie, J. Friedman, & R. Tibshirani (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 371–409). Springer.

Hastie, T., Friedman, J., & Tibshirani, R. (2001b). Support Vector Machines and Flexible Discriminants. In T. Hastie, J. Friedman, & R. Tibshirani (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 371–409). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In *The elements of statistical learning* (pp. 337–387). Springer New York. https://doi.org/10.1007/978-0-387-84858-7_10

Haug, P., Koehler, S., Lau, L. M., Wang, P., Rocha, R., & Huff, S. (1994). A natural language understanding system combining syntactic and semantic techniques. *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 247–251.

Heintzelman, N. H., Taylor, R. J., Simonsen, L., Lustig, R., Anderko, D., Haythornthwaite, J. A., Childs, L. C., & Bova, G. S. (2013). Longitudinal analysis of pain in patients with

metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, *20*(5), 898–905. https://doi.org/10.1136/amiajnl-2012-001076

Hoffmann, D. E., & Tarzian, A. J. (2001). The girl who cried pain: a bias against women in the treatment of pain. *The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics*, *29*(1), 13–27. https://doi.org/10.1111/j.1748-720X.2001.tb00037.x

Horrocks, I. (2005). OWL: A description logic based ontology language. In P. van Beek (Ed.), *Principles and Practice of Constraint Programming - CP 2005* (Vol. 3709, pp. 5–8). Springer Berlin Heidelberg. https://doi.org/10.1007/11564751_2

Horton, D. B., Bhullar, H., Carty, L., Cunningham, F., Ogdie, A., Sultana, J., & Trifirò, G. (2019). Electronic health record databases. In B. L. Strom, S. E. Kimmel, & S. Hennessy (Eds.), *Pharmacoepidemiology* (pp. 241–289). Wiley. https://doi.org/10.1002/9781119413431.ch13

Howard, R., Waljee, J., Brummett, C., Englesbe, M., & Lee, J. (2018). Reduction in Opioid Prescribing Through Evidence-Based Prescribing Guidelines. *JAMA Surgery*, *153*(3), 285–287. https://doi.org/10.1001/jamasurg.2017.4436

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *ArXiv*. https://doi.org/10.48550/arxiv.1904.05342

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., & Wilks, Y. (1998). DESCRIPTION OF THE LaSIE-II SYSTEM AS USED FOR MUC-7 . *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.* Seventh Message Understanding Conference (MUC-7).

Hunston, S. (2002). *Corpora in Applied Linguistics (Cambridge Applied Linguistics)* (1st ed., p. 254). Cambridge University Press.

Husain, M., & Chalder, T. (2021). Medically unexplained symptoms: assessment and management. *Clinical Medicine*, *21*(1), 13–18. https://doi.org/10.7861/clinmed.2020-0947

Hyun, S., Johnson, S. B., & Bakken, S. (2009). Exploring the ability of natural language processing to extract data from nursing narratives. *Computers, Informatics, Nursing : CIN*, *27*(4), 215–223; quiz 224. https://doi.org/10.1097/NCN.0b013e3181a91b58

Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education. (2011). *Relieving pain in america: A blueprint for transforming prevention, care, education, and research*. National Academies Press (US). https://doi.org/10.17226/13172

Institute of Medicine (US) Committee on Data Standards for Patient Safety. (2003). *Key capabilities of an electronic health record system: letter report*. National Academies Press (US). https://doi.org/10.17226/10781

International Association for the Study of Pain. (2005). *Unrelieved Pain is a Major Global Healthcare Problem*.

International Association for the Study of Pain. (2020). *Terminology | International Association for the Study of Pain*. https://www.iasp-pain.org/resources/terminology/

IsHak, W. W., Wen, R. Y., Naghdechi, L., Vanle, B., Dang, J., Knosp, M., Dascal, J., Marcia, L., Gohar, Y., Eskander, L., Yadegar, J., Hanna, S., Sadek, A., Aguilar-Hernandez, L., Danovitch, I., & Louy, C. (2018). Pain and depression: A systematic review. *Harvard Review of Psychiatry*, *26*(6), 352–363. https://doi.org/10.1097/HRP.0000000000000198

I B M. (2021, August 16). *What is Natural Language Processing?* https://www.ibm.com/cloud/learn/natural-language-processing

Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R. J., & Stewart, R. (2017). Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*, *7*(1), e012012. https://doi.org/10.1136/bmjopen-2016-012012

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Resampling Methods. In *An Introduction to Statistical Learning : with Applications in R.* (1st ed.). Springer.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews. Genetics*, *13*(6), 395–405. https://doi.org/10.1038/nrg3208

Jindal, R., & Taneja, S. (2015). A lexical approach for text categorization of medical documents. *Procedia Computer Science*, *46*, 314–320. https://doi.org/10.1016/j.procs.2015.02.026

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural*

Networks and Learning Systems, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Johnson, Alistair, Pollard, Tom, & Mark, Roger. (2015). *MIMIC-III Clinical Database*.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

Johnson, G. J., & Ambrose, P. J. (2006). Neo-tribes. *Communications of the ACM*, *49*(1), 107–113. https://doi.org/10.1145/1107458.1107463

Juckett, D. (2012). A method for determining the number of documents needed for a gold standard corpus. Journal of Biomedical Informatics, 45(3), 460–470. https://doi.org/10.1016/j.jbi.2011.12.010

Jurafsky, Dan, & Martin, J. H. (2023). Transformers and Pretrained Language Models. In *Speech and Language Processing* (3rd (draft)).

Jurafsky, Daniel, & Martin, J. H. (2009). *Speech and Language Processing* (Second).

Kadra, G., Stewart, R., Shetty, H., MacCabe, J. H., Chang, C. K., Taylor, D., & Hayes, R. D. (2018). Long-term antipsychotic polypharmacy prescribing in secondary mental health care and the risk of mortality. *Acta Psychiatrica Scandinavica*, *138*(2), 123–132. https://doi.org/10.1111/acps.12906

Kariotis, T. C., Prictor, M., Chang, S., & Gray, K. (2022). Impact of electronic health records on information practices in mental health contexts: scoping review. *Journal of Medical Internet Research*, *24*(5), e30405. https://doi.org/10.2196/30405

Keshta, I., & Odeh, A. (2020). Security and privacy of electronic health records: Concerns and challenges. *Egyptian Informatics Journal*. https://doi.org/10.1016/j.eij.2020.07.003

Kharrazi, H., Anzaldi, L. J., Hernandez, L., Davison, A., Boyd, C. M., Leff, B., Kimura, J., & Weiner, J. P. (2018). The value of unstructured electronic health record data in geriatric syndrome case identification. *Journal of the American Geriatrics Society*, *66*(8), 1499–1507. https://doi.org/10.1111/jgs.15411

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, *4*, 100057. https://doi.org/10.1016/j.yjbinx.2019.100057

Khoury, M. J., Lam, T. K., Ioannidis, J. P. A., Hartge, P., Spitz, M. R., Buring, J. E., Chanock, S. J., Croyle, R. T., Goddard, K. A., Ginsburg, G. S., Herceg, Z., Hiatt, R. A., Hoover, R. N., Hunter, D. J., Kramer, B. S., Lauer, M. S., Meyerhardt, J. A., Olopade, O. I., Palmer, J. R., … Schully, S. D. (2013). Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers & Prevention*, *22*(4), 508–516. https://doi.org/10.1158/1055-9965.EPI-13-0146

Kim, E., Rubinstein, S. M., Nead, K. T., Wojcieszynski, A. P., Gabriel, P. E., & Warner, J. L. (2019). The evolving use of electronic health records (EHR) for research. *Seminars in Radiation Oncology*, *29*(4), 354–361. https://doi.org/10.1016/j.semradonc.2019.05.010

King, N. B., and Fraser, V. (2013). Untreated pain, narcotics regulation, and global health ideologies. PLoS Medicine, 10(4), e1001411. https://doi.org/10.1371/journal.pmed.1001411

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics*, *84*(11), 956–965. https://doi.org/10.1016/j.ijmedinf.2015.08.004

Koscielny, G., Ison, G., Jupp, S., Parkinson, H., Pendlington, Z. M., Williams, E., Malone, J., Whetzel, T., Sarntivijai, S., Leroy, C., Holloway, E., Adamusiak, T., Hastings, E. K., Ajigboye, O., Roncaglia, P., Kurbatova, N., Welter, D., & Vasant, D. (n.d.). *Experimental Factor Ontology*. Retrieved December 31, 2022, from https://www.ebi.ac.uk/ols/ontologies/efo

Kouiroukidis, N., & Evangelidis, G. (2011). The effects of dimensionality curse in high dimensional knn search. *2011 15th Panhellenic Conference on Informatics*, 41–45. https://doi.org/10.1109/PCI.2011.45

Kraemer, H. C., & Blasey, C. (2016). *How many subjects?: statistical power analysis in research*. SAGE Publications, Ltd. https://doi.org/10.4135/9781483398761

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. https://doi.org/10.1016/j.artmed.2021.102083

Kumar, A., & Wroten, M. (2022). Agnosia. In *StatPearls*. StatPearls Publishing.

Kush, R. D., Helton, E., Rockhold, F. W., & Hardison, C. D. (2008). Electronic health records, medical research, and the Tower of Babel. *The New England Journal of Medicine*, *358*(16), 1738–1740. https://doi.org/10.1056/NEJMsb0800209

K M, A., Basu Roy Chowdhury, S., & Dukkipati, A. (2018). Learning beyond Datasets: Knowledge Graph Augmented Neural Networks for Natural Language Processing. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, *1*, 313–322. https://doi.org/10.18653/v1/N18-1029

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. *ArXiv*. https://doi.org/10.48550/arxiv.1910.09700

Lambert, T. J. R., & Newcomer, J. W. (2009). Are the cardiometabolic complications of schizophrenia still neglected? Barriers to care. *The Medical Journal of Australia*, *190*(S4), S39-42. https://doi.org/10.5694/j.1326-5377.2009.tb02374.x

Landau, S., & Stahl, D. (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research*, *22*(3), 324–345. https://doi.org/10.1177/0962280212439578

Laursen, T. M., Munk-Olsen, T., Agerbo, E., Gasse, C., & Mortensen, P. B. (2009). Somatic hospital contacts, invasive cardiac procedures, and mortality from heart disease in patients with severe mental disorder. *Archives of General Psychiatry*, *66*(7), 713–720. https://doi.org/10.1001/archgenpsychiatry.2009.61

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lehman, L., Saeed, M., Long, W., Lee, J., & Mark, R. (2012). Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annual Symposium Proceedings*, *2012*, 505–511.

Lin, Y., Lu, K., Yu, S., Cai, T., & Zitnik, M. (2023). Multimodal learning on graphs for disease relation extraction. *Journal of Biomedical Informatics*, *143*, 104415. https://doi.org/10.1016/j.jbi.2023.104415

Lin, Z., Yang, D., Jiang, H., & Yin, H. (2021). Learning Patient Similarity via Heterogeneous Medical Knowledge Graph Embedding. *IAENG International Journal of Computer Science*, *48*(4).

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. *ArXiv:2010.11784 [Cs]*.

Liu, K., Hogan, W. R., & Crowley, R. S. (2011). Natural Language Processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, *44*(1), 163–179. https://doi.org/10.1016/j.jbi.2010.07.006

Liu, L., Bustamante, R., Earles, A., Demb, J., Messer, K., & Gupta, S. (2021). A strategy for validation of variables derived from large-scale electronic health record data. *Journal of Biomedical Informatics*, *121*, 103879. https://doi.org/10.1016/j.jbi.2021.103879

Madabushi, H. T., Kochkina, E., & Castelle, M. (2020). *Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data*.

Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *ArXiv*. https://doi.org/10.48550/arxiv.1910.07370

Maniadakis, N., and Gray, A. (2000). The economic burden of back pain in the UK. Pain, 84(1), 95–103. https://doi.org/10.1016/S0304-3959(99)00187-6

Marshall, S. A., Yang, C. C., Ping, Q., Zhao, M., Avis, N. E., & Ip, E. H. (2016). Symptom clusters in women with breast cancer: an analysis of data from social media and a research study. *Quality of Life Research*, *25*(3), 547–557. https://doi.org/10.1007/s11136-015-1156-7

Martin, J. L., McLean, G., Park, J., Martin, D. J., Connolly, M., Mercer, S. W., & Smith, D. J. (2014). Impact of socioeconomic deprivation on rate and cause of death in severe mental illness. *BMC Psychiatry*, *14*(1), 261. https://doi.org/10.1186/s12888-014-0261-4

Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R., & Roberts, A. (2020). Comparative Analysis of Text Classification Approaches in Electronic Health Records. *ArXiv:2005.06624 [Cs]*.

Mascio, A. (2022). *Longitudinal changes in cognitive impairment for patients with Schizophrenia* [Doctoral dissertation, King's College London]. https://kclpure.kcl.ac.uk/portal/en/studentTheses/longitudinal-changes-in-cognitive-impairment-for-patients-with-sc

Mayaud, L., Lai, P. S., Clifford, G. D., Tarassenko, L., Celi, L. A., & Annane, D. (2013). Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. *Critical Care Medicine*, *41*(4), 954–962. https://doi.org/10.1097/CCM.0b013e3182772adb

Ma, R., Perera, G., Romano, E., Vancampfort, D., Koyanagi, A., Stewart, R., Mueller, C., & Stubbs, B. (2022). Predictors of falls and fractures leading to hospitalisation in 36 101 people with affective disorders: a large representative cohort study. *BMJ Open*, *12*(3), e055070. https://doi.org/10.1136/bmjopen-2021-055070

Ma, R., Romano, E., Ashworth, M., Yadegarfar, M. E., Dregan, A., Ronaldson, A., de Oliveira, C., Jacobs, R., Stewart, R., & Stubbs, B. (2022). Multimorbidity clusters among people with serious mental illness: a representative primary and secondary data linkage cohort study. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S003329172200109X

Ma, R., Romano, E., Davis, K., Stewart, R., Ashworth, M., Vancampfort, D., Gaughran, F., Stubbs, B., & Mueller, C. (2022). Osteoporosis referral and treatment among people with severe mental illness: A ten-year data linkage study. *Journal of Psychiatric Research*, *147*, 94–102. https://doi.org/10.1016/j.jpsychires.2022.01.005

McCowan, I. A., Moore, D. C., Nguyen, A. N., Bowman, R. V., Clarke, B. E., Duhig, E. E., & Fry, M.-J. (2007). Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association*, *14*(6), 736–745. https://doi.org/10.1197/jamia.M2130

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Merlin, J. S., Zinski, A., Norton, W. E., Ritchie, C. S., Saag, M. S., Mugavero, M. J., Treisman, G., & Hooten, W. M. (2014). A conceptual framework for understanding chronic pain in patients with HIV. *Pain Practice*, *14*(3), 207–216. https://doi.org/10.1111/papr.12052

Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., & Wong, A. (2021). UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. *ArXiv:2010.10391 [Cs]*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning Based Text Classification: A Comprehensive Review. *ArXiv:2004.03705 [Cs, Stat]*.

Mirza, L., Das-Munshi, J., Chaturvedi, J., Wu, H., Kraljevic, Z., Searle, T., Shaari, S., Mascio, A., Skiada, N., Roberts, A., Bean, D., Stewart, R., Dobson, R., & Bendayan, R. (2021). Investigating the association between physical health comorbidities and disability in individuals with severe mental illness. *European Psychiatry*, *64*(1), e77. https://doi.org/10.1192/j.eurpsy.2021.2255

Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, *22*(2), 1679–1693. https://doi.org/10.1093/bib/bbaa012

Motov, S. M., & Khan, A. N. (2008). Problems and barriers of pain management in the emergency department: Are we ever going to get better? *Journal of Pain Research*, *2*, 5–11.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, *40*(7), 567–576. https://doi.org/10.1080/07357907.2022.2084621

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Naseri, H., Kafi, K., Skamene, S., Tolba, M., Faye, M. D., Ramia, P., Khriguian, J., & Kildea, J. (2021). Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. *Journal of Biomedical Informatics*, *120*, 103864. https://doi.org/10.1016/j.jbi.2021.103864

Naseri, H., Skamene, S., Tolba, M., Faye, M. D., Ramia, P., Khriguian, J., David, M., Kildea, J. (2023). A Scalable Radiomics- and Natural Language Processing–Based Machine Learning Pipeline to Distinguish Between Painful and Painless Thoracic Spinal Bone Metastases: Retrospective Algorithm Development and Validation Study. JMIR AI, 2, e44779. doi: 10.2196/44779

National Library of Medicine. (n.d.). *Use of MeSH in Online Retrieval*. https://www.nlm.nih.gov/mesh/intro_retrieval.html

Negida, A., Fahim, N. K., & Negida, Y. (2019). Sample Size Calculation Guide - Part 4: How to Calculate the Sample Size for a Diagnostic Test Accuracy Study based on Sensitivity, Specificity, and the Area Under the ROC Curve. *Advanced Journal of Emergency Medicine*, *3*(3), e33. https://doi.org/10.22114/ajem.v0i0.158

Neo4j Inc. (2012). *Neo4j - The World's Leading Graph Database*. https://neo4j.com/product/neo4j-graph-database/

NHSBSA. (2023). *Dictionary of medicines and devices (dm+d)* . https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd

NHS Digital. (2022a). *National data opt-out*. NHS Digital. https://digital.nhs.uk/services/national-data-opt-out

NHS Digital. (2022b). *Read Codes*. https://digital.nhs.uk/services/terminology-and-classifications/read-codes

NHS Digital. (2023). *SNOMED CT is an important requirement for electronic patient records*. SNOMED CT. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct#:~:text=The%20use%20of%20SNOMED%20CT,exchanging%20clinical%20information%20between%20systems.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

NIHR Maudsley Biomedical Research Center. (n.d.). *CRIS Data Linkages*. CRIS Data Linkages. Retrieved August 7, 2023, from https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-data-linkages/

NIHR Maudsley Biomedical Research Centre. (n.d.). *Clinical Record Interactive Search (CRIS)*. Retrieved January 11, 2021, from https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/

Noroozian, M., Raeesi, S., Hashemi, R., Khedmat, L., & Vahabi, Z. (2018). Pain: the neglect issue in old people's life. *Open Access Macedonian Journal of Medical Sciences*, *6*(9), 1773–1778. https://doi.org/10.3889/oamjms.2018.335

Nunes, D. A. P., Gomes, J. F., Neto, F., & Matos, D. M. de. (2021). Chronic Pain and Language: A Topic Modelling Approach to Personal Pain Descriptions. *CoRR*, *abs/2109.00402*.

Nuthakki, S., Neela, S., Gichoya, J. W., & Purkayastha, S. (2019). Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks. *ArXiv:1912.12397 [Cs]*.

N H S. (2021a). *Lambeth DataNet*. Lambeth DataNet. https://selondonccg.nhs.uk/wp-content/uploads/2021/04/Lambeth-DataNet.pdf

N H S. (2021b). *Lambeth DataNet*. Lambeth DataNet. https://selondonccg.nhs.uk/wp-content/uploads/2021/04/Lambeth-DataNet.pdf

Newman-Griffis, D., Divita, G., Desmet, B., Zirikly, A., Rosé, C. P., & Fosler-Lussier, E. (2021). Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association*, *28*(3), 516–532. https://doi.org/10.1093/jamia/ocaa269

Onwumere, J., Stubbs, B., Stirling, M., Shiers, D., Gaughran, F., Rice, A. S. C., C de C Williams, A., & Scott, W. (2022). Pain management in people with severe mental illness: an agenda for progress. *Pain*, *163*(9), 1653–1660. https://doi.org/10.1097/j.pain.0000000000002633

Outcomes Based Healthcare. (2018). *Outcomes Platform*. https://outcomesbasedhealthcare.com/outcomes-platform/

Pakhomov, S. S. V., Hemingway, H., Weston, S. A., Jacobsen, S. J., Rodeheffer, R., & Roger, V. L. (2007). Epidemiology of angina pectoris: role of natural language processing of the medical record. *American Heart Journal*, *153*(4), 666–673. https://doi.org/10.1016/j.ahj.2006.12.022

Pakhomov, S. V., Jacobsen, S. J., Chute, C. G., & Roger, V. L. (2008). Agreement between patient-reported symptoms and their documentation in the medical record. *The American Journal of Managed Care*, *14*(8), 530–539.

Pavlou, M., Qu, C., Omar, R. Z., Seaman, S. R., Steyerberg, E. W., White, I. R., & Ambler, G. (2021). Estimation of required sample size for external validation of risk models for binary outcomes. *Statistical Methods in Medical Research*, *30*(10), 2187–2206. https://doi.org/10.1177/09622802211007522

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perera, G., Broadbent, M., Callard, F., Chang, C.-K., Downs, J., Dutta, R., Fernandes, A., Hayes, R. D., Henderson, M., Jackson, R., Jewell, A., Kadra, G., Little, R., Pritchard, M.,

Shetty, H., Tulloch, A., & Stewart, R. (2016). Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open*, *6*(3), e008721. https://doi.org/10.1136/bmjopen-2015-008721

Plaza, L. (2014). Comparing different knowledge sources for the automatic summarization of biomedical literature. *Journal of Biomedical Informatics*, *52*, 319–328. https://doi.org/10.1016/j.jbi.2014.07.014

Polatin, P. B., Kinnedy, R. K., Gatchel, R. J., Lillo, E., & Mayer, T. (1993). Psychiatric Illness and Chronic Low-Back PainThe Mind and the Spine—Which Goes First? *Spine*, *18*(1), 66–71.

Portelli, B., Lenzi, E., Chersoni, E., Serra, G., & Santus, E. (2021). BERT prescriptions to avoid unwanted headaches: A comparison of transformer architectures for adverse drug event detection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1740–1747. https://doi.org/10.18653/v1/2021.eacl-main.149

Potvin, S., & Marchand, S. (2008). Hypoalgesia in schizophrenia is independent of antipsychotic drugs: a systematic quantitative review of experimental studies. *Pain*, *138*(1), 70–78. https://doi.org/10.1016/j.pain.2007.11.007

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W., & Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, *22*(1), 143–154. https://doi.org/10.1136/amiajnl-2013-002544

Public Health England. (2018, September 27). *Severe mental illness (SMI) and physical health inequalities: briefing*. Research and Analysis. https://www.gov.uk/government/publications/severe-mental-illness-smi-physical-health-inequalities/severe-mental-illness-and-physical-health-inequalities-briefing#:~:text=The%20phrase%20severe%20mental%20illness,an%20SMI%20%5Bfootnote%201%5D.

Punchard, N. A., Whelan, C. J., & Adcock, I. (2004). The Journal of Inflammation. *Journal of Inflammation (London, England)*, *1*(1), 1. https://doi.org/10.1186/1476-9255-1-1

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S. (2013). Distributional Semantics Resources for Biomedical Text Processing. *Proceedings of LBM 2013*, 39–44.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *Npj Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1

Rayner, L., Hotopf, M., Petkova, H., Matcham, F., Simpson, A., & McCracken, L. M. (2016). Depression in patients with chronic pain attending a specialised pain treatment centre: prevalence and impact on health care costs. *Pain*, *157*(7), 1472–1479. https://doi.org/10.1097/j.pain.0000000000000542

Research Design Service South Central. (2023). Patient and Public Involvement (PPI). *What Is Patient and Public Involvement in Health and Social Care Research?* https://www.rds-sc.nihr.ac.uk/ppi-information-resources/

Rezaei-Dastjerdehei, M. R., Mijani, A., & Fatemizadeh, E. (2020). Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function. *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, 333–338. https://doi.org/10.1109/ICBME51989.2020.9319440

Richter, A. N., & Khoshgoftaar, T. M. (2019). Learning Curve Estimation with Large Imbalanced Datasets. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 763–768. https://doi.org/10.1109/ICMLA.2019.00135

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical Research Ed.)*, *368*, m441. https://doi.org/10.1136/bmj.m441

Robinson, P. N., & Haendel, M. A. (2020). Ontologies, knowledge representation, and machine learning for translational research: recent contributions. *Yearbook of Medical Informatics*, *29*(1), 159–162. https://doi.org/10.1055/s-0040-1701991

Roesslein, J. (2020). *Tweepy: Twitter for Python!* https://github.com/tweepy/tweepy

Rowlands, G., Whitney, D., & Moon, G. (2018). Developing and applying geographical synthetic estimates of health literacy in GP clinical systems. *International Journal of Environmental Research and Public Health*, *15*(8). https://doi.org/10.3390/ijerph15081709

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., & Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical Care Medicine*, *39*(5), 952–960. https://doi.org/10.1097/CCM.0b013e31820a92c6

Sager, N., Lyman, M., Tick, L. J., Nhàn, N. T., & Bucknall, C. E. (1993). Natural language processing of asthma discharge summaries for the monitoring of patient care. *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 265–268.

Sammut, C., & Webb, G. I. (Eds.). (2010). Logistic Regression. In *Encyclopedia of Machine Learning* (pp. 631–631). Springer US.

Sampson, R., Cooper, J., Barbour, R., Polson, R., & Wilson, P. (2015). Patients' perspectives on the medical primary-secondary care interface: systematic review and synthesis of qualitative research. *BMJ Open*, *5*(10), e008708. https://doi.org/10.1136/bmjopen-2015-008708

Samulowitz, A., Gremyr, I., Eriksson, E., & Hensing, G. (2018). "Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Research & Management : The Journal of the Canadian Pain Society = Journal de La Societe Canadienne Pour Le Traitement de La Douleur*, *2018*, 6358624. https://doi.org/10.1155/2018/6358624

Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Strauss, M., Geyer, P. E., Coscia, F., Albrechtsen, N. J. W., Mundt, F., Jensen, L. J., & Mann, M. (2022). A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology*, *40*(5), 692–702. https://doi.org/10.1038/s41587-021-01145-6

Sastre, J., Zaman, F., Duggan, N., McDonagh, C., & Walsh, P. (2020). A deep learning knowledge graph approach to drug labelling. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2513–2521. https://doi.org/10.1109/BIBM49941.2020.9313350

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. *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*, 139–144. https://doi.org/10.18653/v1/D19-3024

Shai, B.-D. (2014). Stochastic Gradient Descent. In S.-S. Shai (Trans.), *Understanding Machine Learning: From Theory to Algorithms* (pp. 150–166). Cambridge University Press.

Sharma, R., Wigginton, B., Meurk, C., Ford, P., & Gartner, C. E. (2016). Motivations and Limitations Associated with Vaping among People with Mental Illness: A Qualitative Analysis of Reddit Discussions. *International Journal of Environmental Research and Public Health*, *14*(1). https://doi.org/10.3390/ijerph14010007

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, *21*(2), 221–230. https://doi.org/10.1136/amiajnl-2013-001935

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*(11), 1297–1304. https://doi.org/10.1093/jamia/ocz096

Smith, B., Ceusters, W., Goldberg, L. J., & Ohrbach, R. (2011). Towards an Ontology of Pain. *Proceedings of the Conference on Ontology and AnalyticalMetaphysics* , 23–36.

Smyth, C. (2010). An introduction to corpus linguistics. *Journal of English Linguistics*, *28*(2), 193–196. https://doi.org/10.1177/00754240022004965

Snijders, R. A. H., Brom, L., Theunissen, M., & van den Beuken-van Everdingen, M. H. J. (2023). Update on Prevalence of Pain in Patients with Cancer 2022: A Systematic Literature Review and Meta-Analysis. *Cancers*, *15*(3). https://doi.org/10.3390/cancers15030591

Social Media Today. (n.d.). *Reddit Now Has as Many Users as Twitter, and Far Higher Engagement Rates*. Social Media Today. Retrieved March 8, 2021, from https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/

Sordo, M., & Zeng, Q. (2005). On sample size and classification accuracy: A performance comparison. In J. L. Oliveira, V. Maojo, F. Martín-Sánchez, & A. S. Pereira (Eds.), *Biological and medical data analysis* (Vol. 3745, pp. 193–201). Springer Berlin Heidelberg. https://doi.org/10.1007/11573067_20

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.

Stephens, J., Laskin, B., Pashos, C., Peña, B., and Wong, J. (2003). The burden of acute postoperative pain and the potential role of the COX-2-specific inhibitors. Rheumatology (Oxford, England), 42 Suppl 3, iii40-52. https://doi.org/10.1093/rheumatology/keg497

Stewart, R., & Davis, K. (2016). "Big data" in mental health research: current status and emerging possibilities. *Social Psychiatry and Psychiatric Epidemiology*, *51*(8), 1055–1072. https://doi.org/10.1007/s00127-016-1266-8

Stewart, R., Soremekun, M., Perera, G., Broadbent, M., Callard, F., Denis, M., Hotopf, M., Thornicroft, G., & Lovestone, S. (2009). The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry*, *9*, 51. https://doi.org/10.1186/1471-244X-9-51

Strassnig, M., Brar, J. S., & Ganguli, R. (2003). Body mass index and quality of life in community-dwelling patients with schizophrenia. *Schizophrenia Research*, *62*(1–2), 73–76. https://doi.org/10.1016/s0920-9964(02)00441-3

Stubbs, A., Filannino, M., Soysal, E., Henry, S., & Uzuner, Ö. (2019). Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, *26*(11), 1163–1171. https://doi.org/10.1093/jamia/ocz163

Stubbs, A., & Uzuner, Ö. (2019). New approaches to cohort selection. *Journal of the American Medical Informatics Association*, *26*(11), 1161–1162. https://doi.org/10.1093/jamia/ocz174

Stubbs, B, Eggermont, L., Mitchell, A. J., De Hert, M., Correll, C. U., Soundy, A., Rosenbaum, S., & Vancampfort, D. (2015). The prevalence of pain in bipolar disorder: a systematic review and large-scale meta-analysis. *Acta Psychiatrica Scandinavica*, *131*(2), 75–88. https://doi.org/10.1111/acps.12325

Stubbs, Brendon, Gardner-Sood, P., Smith, S., Ismail, K., Greenwood, K., Patel, A., Farmer, R., & Gaughran, F. (2015). Pain is independently associated with reduced health related quality of life in people with psychosis. *Psychiatry Research*, *230*(2), 585–591. https://doi.org/10.1016/j.psychres.2015.10.008

Stubbs, Brendon, Mitchell, A. J., De Hert, M., Correll, C. U., Soundy, A., Stroobants, M., & Vancampfort, D. (2014). The prevalence and moderators of clinical pain in people with schizophrenia: a systematic review and large scale meta-analysis. *Schizophrenia Research*, *160*(1–3), 1–8. https://doi.org/10.1016/j.schres.2014.10.017

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, *20*(5), 806–813. https://doi.org/10.1136/amiajnl-2013-001628

Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *ArXiv*. https://doi.org/10.48550/arxiv.1902.10197

Tafvizi, A., Avci, B., & Sundararajan, M. (2022). *Attributing AUC-ROC to Analyze Binary Classifier Performance*.

Tan, W. K., Hassanpour, S., Heagerty, P. J., Rundell, S. D., Suri, P., Huhdanpaa, H. T., James, K., Carrell, D. S., Langlotz, C. P., Organ, N. L., Meier, E. N., Sherman, K. J., Kallmes, D. F., Luetmer, P. H., Griffith, B., Nerenz, D. R., & Jarvik, J. G. (2018).

Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain. *Academic Radiology*, *25*(11), 1422–1432. https://doi.org/10.1016/j.acra.2018.03.008

Taylor, P. (2008). Personal genomes: when consent gets in the way. *Nature*, *456*(7218), 32–33. https://doi.org/10.1038/456032a

Thompson, T., Correll, C. U., Gallop, K., Vancampfort, D., & Stubbs, B. (2016). Is Pain Perception Altered in People With Depression? A Systematic Review and Meta-Analysis of Experimental Pain Research. *The Journal of Pain*, *17*(12), 1257–1272. https://doi.org/10.1016/j.jpain.2016.08.007

Tian, T. Y., Zlateva, I., & Anderson, D. R. (2013). Using electronic health records data to identify patients with chronic pain in a primary care setting. *Journal of the American Medical Informatics Association*, *20*(e2), e275-80. https://doi.org/10.1136/amiajnl-2013-001856

Tracey, I. (2016). Finding the hurt in pain. *Cerebrum : The Dana Forum on Brain Science*, *2016*.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. *ArXiv*. https://doi.org/10.48550/arxiv.1606.06357

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

Twitter Help Center. (n.d.). *About Twitter's APIs*. Retrieved March 16, 2021, from https://help.twitter.com/en/rules-and-policies/twitter-api

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256. https://doi.org/10.1038/s41598-022-10358-x

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., & South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, *19*(5), 786–791. https://doi.org/10.1136/amiajnl-2011-000784

Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, *15*(1), 14–24. https://doi.org/10.1197/jamia.M2408

Uzuner, O., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, *14*(5), 550–563. https://doi.org/10.1197/jamia.M2444

Uzuner, O., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, *17*(5), 514–518. https://doi.org/10.1136/jamia.2010.003947

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203

Uzuner, Ö., Stubbs, A., & Filannino, M. (2017). A natural language processing challenge for clinical records: Research Domains Criteria (RDoC) for psychiatry. *Journal of Biomedical Informatics*, *75S*, S1–S3. https://doi.org/10.1016/j.jbi.2017.10.005

Uzuner, Ö., Stubbs, A., & Lenert, L. (2020). Advancing the state of the art in automatic extraction of adverse drug events from narratives. *Journal of the American Medical Informatics Association*, *27*(1), 1–2. https://doi.org/10.1093/jamia/ocz206

Uzuner, Ö., & Stubbs, A. (2015). Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. *Journal of Biomedical Informatics*, *58 Suppl*(Suppl), S1–S5. https://doi.org/10.1016/j.jbi.2015.10.007

Uzuner, O. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, *16*(4), 561–570. https://doi.org/10.1197/jamia.M3115

Vallerand, A. H., & Polomano, R. C. (2000). The relationship of gender to pain. *Pain Management Nursing : Official Journal of the American Society of Pain Management Nurses*, *1*(3 Suppl 1), 8–15. https://doi.org/10.1053/jpmn.2000.9759

Vandenbussche, N., Van Hee, C., Hoste, V., & Paemeleire, K. (2022). Using natural language processing to automatically classify written self-reported narratives by patients with migraine or cluster headache. *The Journal of Headache and Pain*, *23*(1), 129. https://doi.org/10.1186/s10194-022-01490-0

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, *18*(1), 148. https://doi.org/10.1186/s12874-018-0594-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv*. https://doi.org/10.48550/arxiv.1706.03762

Velupillai, S, Mowery, D., South, B. R., Kvist, M., & Dalianis, H. (2015). Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of Medical Informatics*, *10*(1), 183–193. https://doi.org/10.15265/IY-2015-009

Velupillai, Sumithra, Mowery, D. L., Conway, M., Hurdle, J., & Kious, B. (2016). Vocabulary Development To Support Information Extraction of Substance Abuse from Psychiatry Notes. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 92–101. https://doi.org/10.18653/v1/W16-2912

Velupillai, Sumithra, Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, *88*, 11–19. https://doi.org/10.1016/j.jbi.2018.10.005

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, *58*(5), 475–483. https://doi.org/10.1016/j.jclinepi.2004.06.017

Viana, M. C., Lim, C. C. W., Garcia Pereira, F., Aguilar-Gaxiola, S., Alonso, J., Bruffaerts, R., de Jonge, P., Caldas-de-Almeida, J. M., O'Neill, S., Stein, D. J., Al-Hamzawi, A., Benjet, C., Cardoso, G., Florescu, S., de Girolamo, G., Haro, J. M., Hu, C., Kovess-Masfety, V., Levinson, D., … Scott, K. M. (2018). Previous mental disorders and

subsequent onset of chronic back or neck pain: findings from 19 countries. *The Journal of Pain*, *19*(1), 99–110. https://doi.org/10.1016/j.jpain.2017.08.011

Viani, N., Botelle, R., Kerwin, J., Yin, L., Patel, R., Stewart, R., & Velupillai, S. (2021). A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, *11*(1), 757. https://doi.org/10.1038/s41598-020-80457-0

Viani, N., Kam, J., Yin, L., Bittar, A., Dutta, R., Patel, R., Stewart, R., & Velupillai, S. (2020). Temporal information extraction from mental health records to identify duration of untreated psychosis. *Journal of Biomedical Semantics*, *11*(1), 2. https://doi.org/10.1186/s13326-020-00220-2

Viani, N., Patel, R., Stewart, R., & Velupillai, S. (2019). Generating Positive Psychosis Symptom Keywords from Electronic Health Records. In D. Riaño, S. Wilk, & A. ten Teije (Eds.), *Artificial intelligence in medicine* (Vol. 11526, pp. 298–303). Springer International Publishing. https://doi.org/10.1007/978-3-030-21642-9_38

Vinall, J., Pavlova, M., Asmundson, G. J. G., Rasic, N., & Noel, M. (2016). Mental health comorbidities in pediatric chronic pain: A narrative review of epidemiology, models, neurobiological mechanisms and treatment. *Children (Basel, Switzerland)*, *3*(4). https://doi.org/10.3390/children3040040

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & STROBE Initiative. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Annals of Internal Medicine*, *147*(8), 573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010

Von Korff, M., DeBar, L. L., Deyo, R. A., Mayhew, M., Kerns, R. D., Goulet, J. L., & Brandt, C. (2020). Identifying Multisite Chronic Pain with Electronic Health Records Data. *Pain Medicine*, *21*(12), 3387–3392. https://doi.org/10.1093/pm/pnaa295

Wandner, L. D., Scipio, C. D., Hirsh, A. T., Torres, C. A., & Robinson, M. E. (2012). The perception of pain in others: how gender, race, and age influence pain expectations. *The Journal of Pain*, *13*(3), 220–227. https://doi.org/10.1016/j.jpain.2011.10.014

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, *29*(12), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20. https://doi.org/10.1016/j.jbi.2018.09.008

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Webb, G. I., Keogh, E., Miikkulainen, R., Miikkulainen, R., & Sebag, M. (2010). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 713–714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

Weber, G. M., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. *The Journal of the American Medical Association*, *311*(24), 2479–2480. https://doi.org/10.1001/jama.2014.4228

Weng, Y., Tian, L., Tedesco, D., Desai, K., Asch, S. M., Carroll, I., Curtin, C., McDonald, K. M., & Hernandez-Boussard, T. (2020). Trajectory analysis for postoperative pain using electronic health records: A nonparametric method with robust linear regression and K-medians cluster analysis. *Health Informatics Journal*, *26*(2), 1404–1418. https://doi.org/10.1177/1460458219881339

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Widnall, E., Epstein, S., Polling, C., Velupillai, S., Jewell, A., Dutta, R., Simonoff, E., Stewart, R., Gilbert, R., Ford, T., Hotopf, M., Hayes, R. D., & Downs, J. (2022). Autism spectrum disorders as a risk factor for adolescent self-harm: a retrospective cohort study of 113,286 young people in the UK. *BMC Medicine*, *20*(1), 137. https://doi.org/10.1186/s12916-022-02329-w

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., … Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*. https://doi.org/10.48550/arxiv.1910.03771

Woodhead, C., Ashworth, M., Schofield, P., & Henderson, M. (2014). Patterns of physical co-/multi-morbidity among patients with serious mental illness: a London borough-based cross-sectional study. *BMC Family Practice*, *15*, 117. https://doi.org/10.1186/1471-2296-15-117

World Health Organization. (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision* (2nd ed, p. Spanish version, 1st edition published by PAHO as Publicación Científica 544). World Health Organization.

World Health Organization. (2008). *International Classification of Diseases (ICD)*. https://www.who.int/standards/classifications/classification-of-diseases

Wright, A., McCoy, A. B., Henkin, S., Kale, A., & Sittig, D. F. (2013). Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. *Journal of the American Medical Informatics Association*, *20*(5), 887–890. https://doi.org/10.1136/amiajnl-2012-001576

Wu, Y., Liu, Z., Wu, L., Chen, M., & Tong, W. (2021). BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study for Classifying Drug-Induced Liver Injury Risk. *Frontiers in Artificial Intelligence*, *4*, 729834. https://doi.org/10.3389/frai.2021.729834

Xing, W., & Bei, Y. (2020). Medical health big data classification based on KNN classification algorithm. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754

Yan, D., and Guo, S. (2019). Leveraging contextual sentences for text classification by using a neural attention model. Computational Intelligence and Neuroscience, 2019, 8320316. https://doi.org/10.1155/2019/8320316

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, *19*(Suppl 3), 71. https://doi.org/10.1186/s12911-019-0781-4

Ye, C., & Fabbri, D. (2018). Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *Journal of Biomedical Informatics*, *83*, 63–72. https://doi.org/10.1016/j.jbi.2018.05.014

Ye, Q., Yang, R., Cheng, C.-L., Peng, L., & Lan, Y. (2023). Combining the external medical knowledge graph embedding to improve the performance of syndrome differentiation model. *Evidence-Based Complementary and Alternative Medicine*, *2023*, 2088698. https://doi.org/10.1155/2023/2088698

Yim, W., Yetisgen, M., Harris, W., & Kwan, S. (2016). Natural Language Processing in Oncology: A Review | Electronic Health Records | JAMA Oncology | JAMA Network. *JAMA Oncology*.

Yoon, B.-H., Kim, S.-K., & Kim, S.-Y. (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics & Informatics*, *15*(1), 19–27. https://doi.org/10.5808/GI.2017.15.1.19

Zou, Y., Pesaranghader, A., Song, Z., Verma, A., Buckeridge, D. L., & Li, Y. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Scientific Reports*, *12*(1), 17868. https://doi.org/10.1038/s41598-022-22956-w