NHS England

Internship Scheme for Innovation and Analytics in Health

Technical Project Report

Enriching Neurology Patient Information Using MedCAT:

Evaluation of Pre-trained Named Entity Recognition & Linking Models

Data Science Intern
Nur Aizaan Anwar
PhD Candidate in AI & Machine Learning,
UKRI AI4Health Centre for Doctoral Training,
Imperial College London

Supervisor
Paul Carroll
Senior Data Scientist,
Digital Analytics and Research Team,
NHS England

Supervisor
Jonathan Pearson
Lead Data Scientist,
Digital Analytics and Research Team,
NHS England

Collaborator
Vishnu Chandrabalan
Head of Data Science & AI,
Lancashire Teaching Hospitals NHS Foundation Trust

Collaborator
Joanne Knight
Chair in Applied Data Science,
Centre for Health Informatics, Computing & Statistics,
Lancaster University

Collaborator
Hedley Emsley
Consultant Neurologist, Lancashire Teaching Hospitals NHS Foundation Trust
&
Professor of Clinical Neuroscience,
Lancaster University

September 14, 2023

Abstract

Neurology and other clinical specialities are awash with clinical data. However, these are generally not structured and lack the characteristics to allow straightforward automatic extraction of clinically relevant concepts. Software tools do exist that can recognise clinical terms in unstructured clinical data (e.g. clinic letters) and link them to other concepts. These are called 'named entity recognition and linking' (NER+L) tools. But many such tools require prior 'labelling' by a domain expert (i.e. person with specialty knowledge) of the relevant clinical concepts. MedCAT is a NER+L tool that can work without this prior labelling as it contains an algorithm that is aligned with a customisable knowledge database (ontology). This works in two stages: 1) linking unambiguous portions of texts (entities) to unique terms in the ontology then 2) linking ambiguous entities to terms in the ontology with the most similar contexts. However, evaluation of the MedCAT models which inform the NER+L process has only been performed on labelled data, and the learned numerical representations of concepts (embeddings) has not been assessed before. Additionally, to the best of our knowledge, in the UK, only organisations related to the home institution (King's College) have evaluated MedCAT. We are the first NHS organisation, independent of King's College, to have evaluated MedCAT models and its usability. The contributions of this report are: 1) evaluation of three separate MedCAT models, 2) comparison of three different clustering techniques as evaluation methods in the absence of labelled data, 3) evaluation of MedCAT's learned concept embeddings, 4) comparison of intrinsic and extrinsic evaluation metrics and 5) comparison of qualitative and quantitative evaluation approaches. We found that all three models produced NER+L results which are not consistent with clinical understanding. Clustering can enable deeper examination of learned embeddings, but further work needs to be done on finding the best input data and clustering approach. Intrinsic evaluation metrics are only meaningful in the presence of extrinsic measures and further research needs to be done to identify the most informative set of metrics. Quantitative assessment must be supplemented by qualitative inspection. The work performed here forms the first phase in evaluation of MedCAT models' performance. Once optimal evaluation strategies have been identified, the next phase can be focused on improving MedCAT models. This will ultimately enable extraction of clinical terms that can be used for multiple downstream tasks such as automated clinical coding, research, monitoring of interventions, audits as well as service improvements.

Contents

List of Tables

List of Figures

Chapter 1

Introduction & Background

1.1 Introduction

1.1.1 The Need for Automated Clinical Coding in Neurology in the UK

Neurology, like any medical speciality, contains a wealth of clinical information waiting to be harnessed[1,2]. The advent of electronic health records shows great promise in enabling access to this resource[1]. However, rather than structured fields, much of clinical data is contained in free text, which, in its raw form cannot be analysed[3]. Whilst it may be more convenient for an end-user to express the complexity of patient care in an unstructured manner, it poses a challenge for health services to gain insight from a monolithic form of health data[3]. Transforming unstructured clinical data into a structured form can be achieved via clinical coding. This process involves assigning standardised codes to free text, obtained from a classification system (ontology) such as the Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT)[4,5].

Clinical coding in the UK is performed manually by non-medical staff who have been trained to analyse, summarise and classify clinical texts in a standardised manner[6]. Manual coding, although can achieve an accuracy of greater than 90%, is time-consuming, error-prone, inconsistent and labour-intensive[4,7–10]. Despite the ability to code thousands of health records per month, there still exists huge backlogs of cases for UK National Health Service (NHS) coding departments, which may take over a year to clear[7]. This issue is compounded by the fact that diagnostic coding is not mandated in the outpatient setting, where much of neurology care is delivered[6]. As such, there is no consistent practice in neurology outpatient coding, and there exists variations in systems amongst organisations and individuals. As an oversubscribed and under-served health service[11,12], neurology outpatient care can benefit immensely from coding of free text. Structured data can improve understanding of the state of service; patient diagnoses and treatments; resource allocation; and performance of interventions, be they at the patient- or population-level[1,3,6,13]. Automating clinical coding with computers and artificial intelligence (AI), can potentially expedite this activity and alleviate the burden on manual coders. This process is known as Automated Clinical Coding (ACC)[4].

1.1.2 Project Scope & Goal

The Medical Concept Annotation Toolkit (MedCAT) is a software package developed by King's College London (KCL), that can train a mathematical model to extract clinically-relevant concepts and link them to a chosen ontology[14]. This approach, called Named Entity Recognition and Linking (NER+L), can be used for development of ACC. However, in the UK, the performance of these models has yet to be assessed independent of its home institution of KCL. In addition, all evaluations of MedCAT's models' performance have been done using labelled data and the assessment of the numerical representations (embeddings) of linked concepts have not been performed before[14–22].

In collaboration with Lancashire Teaching Hospitals NHS Foundation Trust (LTH), we present early work on evaluating MedCAT models in the absence of human-annotated labels. The scope of our project is limited to the concepts which are classed by SNOMED CT as 'Disorders'. The rationale for this is that our stakeholders aim to create a database of medical documents that is searchable based on medical conditions. SNOMED CT is the coding system chosen for this work as it has been found to be user-friendly and extensive enough to capture the complexity of neurological conditions[23]. Furthermore, using one type of concept also allows for ease of inspection of the NER+L results and MedCAT models. In this work, we neither train MedCAT NER+L models from scratch, nor do we further train pre-trained models with LTH data. Instead, we assess the NER+L performed by three pre-trained models on LTH data, then examine the NER+L models themselves qualitatively and quantitatively.

The research questions we hope to address in this work are as follows:

1. What is the NER+L frequency of MedCAT models (coverage) when applied to LTH data and how do they compare with each other?
2. To what extent has the MedCAT models learned?
3. How do MedCAT models compare with each other?
4. How do our clustering methods compare with each other?
5. How do qualitative and quantitative evaluation approaches compare with each other?
6. How do intrinsic and extrinsic metrics compare with each other?

Section 1.2 (Background) presents the nature of unstructured clinical data along with their analytical challenges; an overview of ACC methods; NER+L as an ACC approach; MedCAT as a NER+L tool; and evaluations already performed on MedCAT. Chapter 2 (Methods) outlines the approach taken to evaluate three MedCAT models. Chapter 3 (Results) presents main findings from application of MedCAT models to our dataset along with early NER+L evaluation. Chapter 4 (Discussion) presents a summary of key findings and their implications; limitations of our approach; and suggestions for future work. Finally, Chapter 5 (Conclusions) summarises key messages from this work.

1.2 Background

1.2.1 Analytical Challenges of Unstructured Clinical Records

To be effective at ACC, computer systems must have the ability to handle the idiosyncratic nature of medical records such as usage of abbreviations (e.g. BIBA, h/o, HPC, etc); inconsistent document lengths and layouts; special characters (e.g. ♂ for male and ♀ for female); context (e.g. negation, temporality of conditions and experiencing subject); misspellings; redundant text; ambiguous terminology; and grammatical errors[4,9,24,25]. They must also be able to adapt to the unique documentation styles of every clinical speciality, healthcare organisation and medical practitioner. Development of ACCs must also be done in view of limited training data availability (given medico-legal restrictions), poor data quality and lack of gold standard labels[1–4,7,24–26].

1.2.2 Overview of Automated Clinical Coding Methods

Broadly, ACC can be performed either via manually pre-specified rules (rule-based techniques) or automatically learned rules i.e. machine learning (ML) methods based on analysis of human language (Natural Language Processing, or NLP)[3,13,27,28]. Rule-based techniques, although having been demonstrated to yield high positive predictive value (precision), suffer from poor sensitivity (recall) and requires explicit clinician-informed rules[4,13,29]. Rule-writing becomes increasingly complex and laborious in order to match the infinite variations in medical text to the tens-of-thousands of ever-expanding clinical codes[9,10]. NLP methods can circumvent the challenges of rule-based techniques through learning of the text-to-code rules implicitly[13]. This can be achieved either via learning human-labelled annotations (supervised learning) and/or inherent arrangements in the data itself (unsupervised learning)[29]. Supervised methods frequently demonstrate higher accuracy, yet are often constrained by training data availability and biases in coding practices[30]. However, ML algorithms need not necessarily be used in isolation, and the parameters learned from one model training can be used in another in a process called 'transfer learning'[31]. Model 'fine-tuning' occurs when the parameters are updated for a customised (downstream) task[31–33]. The choice of using either raw text or pre-extracted information from text as inputs (features) to an algorithm depends on the ML technique used and the desired outcome[9,25,29].

1.2.3 Named Entity Recognition and Linking as a Route to Automated Clinical Coding

A type of feature that can be used by computer algorithms to learn clinical codes is called 'named entities'[4,30,34–37]. Named entities are portions of texts that represent a concept which can be grouped with other concepts of similar meaning to form a category (semantic type)[34,38–43]. For example, 'diabetic neuropathy' is a portion of text that can be represented by the concept 'peripheral nerve disease', which falls under the semantic type of 'disorder'. The process of identifying named entities is known as Named Entity Recognition (NER)[42–44]. When identified entities are linked to concepts in a reference database i.e. knowledge base, ontology or classification system, such as the International Classification of Diseases, Tenth Revision (ICD-10)[45]; the Unified Medical Language System (UMLS)[46]; and SNOMED CT[5], the process is known as 'linking'[47–52]. NER, coupled with linking (NER+L) allows for standardisation, transferability, categorisation and comparability of health records[53].

1.2.4 Medical Concept Annotation Toolkit (MedCAT) as a Named Entity Recognition and Linking Tool

The Medical Concept Annotation Toolkit (MedCAT) is an example of a NER+L software which has been developed to support analysis of unstructured clinical data[14]. It is an open-source suite of tools for NLP pre-processing, ML, NER+L, contextualisation, inspection and annotation[14] . Figure 1.1 outlines the workflow of the MedCAT technology. ML training and NER+L begins with a vocabulary and a concept database (CDB). The vocabulary contains all words that can be annotated and their corresponding vector representations. For our work, we do not build the vocabulary or CDB from scratch, but use the versions compiled by the MedCAT team (see the script to build the Vocabulary here: https://github.com/CogStack/MedCAT/blob/master/medcat/vocab.py). The vector representations of those words were obtained by training 300-dimensional Word2Vec embeddings[54] on the Medical Information Mart for Intensive Care (MIMIC-III) dataset[55].[14] An example of a vocabulary of two words, followed by the count for the word in the dataset and the corresponding 3-dimensional word embedding is given in Figure

1.2. The CDB is a customisable table of concepts from one or more ontologies and all of its synonyms (names). Given the vocabulary, an unstructured corpus is pre-processed, spell-checked, then tokenised and lemmatised with the biomedical NLP library, SciSpacy[14,56].

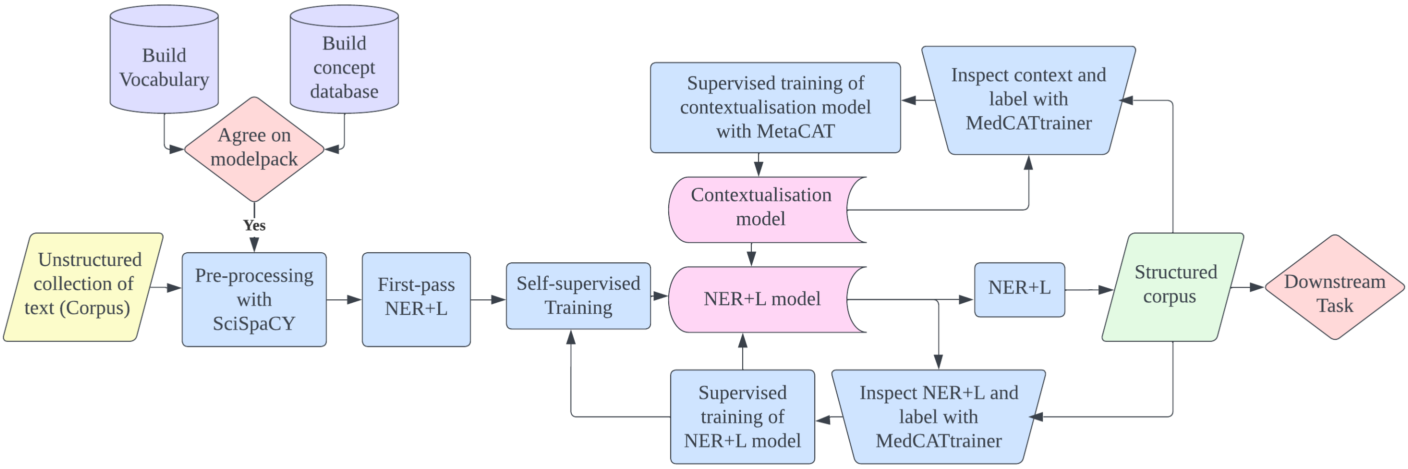
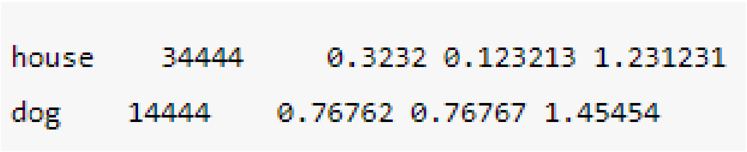

Build Vocabulary
Build concept database
Agree on modelpack
Yes
Unstructured collection of text (Corpus)
Pre-processing with SciSpaCY
First-pass NER+L
Self-supervised Training
Supervised training of contextualisation model with MetaCAT
Inspect context and label with MedCATtrainer
Contextualisation model
NER+L model
NER+L
Structured corpus
Downstream Task
Supervised training of NER+L model
Inspect NER+L and label with MedCATtrainer

Figure 1.1: Schematic representation of the MedCAT workflow summarised from Kraljevic et al 2021[14].



```
house    34444     0.3232 0.123213 1.231231
dog    14444    0.76762 0.76767 1.45454
```

Figure 1.2: Example vocabulary of two words with 3-dimensional word embeddings. The leftmost column lists the words that appear in a corpus. The centre column gives the count of the word in the dataset. The rightmost column lists three numbers which represent 3-dimensions of word embeddings that represent the word in the leftmost column.[57]

First-pass NER+L, Self-supervised Training & NER+L model

Following pre-processing, the first-pass NER+L algorithm detects entity candidates by marking terms in the corpus which are concepts in the CDB. The context vectors of these detected entities are then learned (see Equation 2 in Kralkevic et al)[14]. If the entity is unique (maps one-to-one to a CDB concept), it is linked to its corresponding concept and that concept's embedding is updated (see Equation 10 in Kraljevic et al)[14]. For non-unique (ambiguous) entities, they require disambiguation. This is done by calculating the similarity between their context vectors and unique entities' context vectors. The ambiguous entity is linked to a concept whose similarity exceeds a pre-set threshold. The concept embedding is then updated. At the end of training, concepts linked to detected entities would have 300-dimensional embeddings that are reflections of the contexts from which they were updated. Only concept embeddings are stored and can be accessed by the end-user via the `cat.cdb.cui2context_vectors['concept unique identifier']` Python dictionary method. There would be multiple types of the learned concept embedding per concept depending on the configurations of the context window of a detected entity. These embeddings represent the NER+L model.

Supervised Training

For supervised training (finetuning), labels (correct concepts) would be provided to selected entities via manual annotation performed on the MedCATtrainer interface. The context embeddings of these entities are calculated and their corresponding concept embeddings updated.[14].

MetaCAT for annotation of context

MetaCAT is a separate algorithm that uses the context of MedCAT entities in a multiclass classification task using a bilateral Long-Short-Term Memory Network[14]. It requires labelling of concepts by domain experts. The goal is to label entities according to their context (in a process known as meta-annotation or contextualisation). These contexts can include attributes such as negation, presence, temporality and experiencer. The MetaCAT model can be used in conjuction with the NER+L function or separately.

NER+L

Using the learned concept embeddings (NER+L model) with/without a MetaCAT model, NER+L (annotation) is performed as per the first-pass NER+L and disambiguation process.

Structured Annotation Corpus

Following NER+L, a structured annotation corpus is produced in the form of Python nested dictionaries. The most superficial key in the dictionary represents the document identifier, which corresponds to the index of the input document. Each document identifier would contain Python dictionaries of entities that have been linked to the CDB. Each entity would contain a Python dictionary of the associated concept name, semantic type (if applicable), context similarity and contextualised task result (if applicable). The structured annotation corpus can be used for further downstream tasks such as clinical coding.

1.2.5 Reported Evaluation of MedCAT

The reported strength of MedCAT is that the majority of its NER+L is performed using a self-supervised ML algorithm (an unsupervised ML method where the algorithm learns labels from its own data)[14]. Inspection of NER+L results is performed manually using a separate interface (MedCATtrainer) which also allows for human annotation, contextualisation and supervised training. NER+L models were initially trained and finetuned at MedCAT's home clinical institution, King's College Hospital Foundation Trust (KCH), then further trained at partner institutions: 1) South London and Maudsley Foundation Trust and 2) University College London Hospitals Foundation Trust. To the best of our knowledge, in the UK, MedCAT has only been validated on real-world data from clinical centres related to the home institution of KCL, using models trained initially on the freely-available MIMIC-III dataset[14–18,22,55].

Three other studies which evaluated MedCAT's performance were conducted outside the UK, and the results were heterogenous[19–21]. Using transfer learning of a publicly-available MedCAT model, Ariño et al found a NER+L (annotation) validation accuracy of 86.7% for MedCAT's annotation of the MIMIC-III dataset[21]. Although the validation accuracy was deemed good by the authors, it is no surprise that this was achieved, as the public MedCAT model was trained on MIMIC-III itself, and the validation may have actually been done on 'seen' data. Kunz et al evaluated NER+L with MedCAT using a text dataset (corpus) of 45 documents translated automatically from German to English[19]. Five documents were manually annotated. They conducted three NER+L scenarios using a SNOMED CT-based model: 1) untrained, 2) trained self-supervised and 3) trained supervised (using three out of five annotated documents for training). The harmonic mean of precision and recall (F1 measure) was 0.5 for all scenarios and this is noticeably worse than the NER+L F1 result of the base MedCAT model achieved by the original developers (0.638)[14]. It is however, challenging to directly compare these results as Kunz et al did not specify the MedCAT model used in transfer learning[19]. Finally, van Es et al only evaluated MetaCAT's supervised ML algorithm's performance at negation detection in Dutch clinical notes against a rule-based method and another supervised algorithm (finetuned RoBERTa)[20]. Both ML methods outperformed the rule-based approach, with the fintetuned RoBERTa being the best overall. In summary, all evaluation performed on MedCAT NER+L so far have been validated on manually-labelled data which requires expert input (domain-expertise) and is labour-intensive to obtain. This creates an opportunity not only for MedCAT's NER+L to be evaluated on unannotated data, but also in an NHS organisation that is independent of the home institution.

Chapter 2

Methods

All analytical work has been performed on the remote Data Science Virtual Desktop based in the Lancashire Data Science Environment (LANDER)[58].

2.1 Project Scoping

Project scoping was performed in collaboration with colleagues from LTH. We identified that the outpatient neurology service would benefit from a database of medical documents searchable by SNOMED CT clinical conditions. Before this can be achieved, we must ensure that the model used for NER+L (updated/ learned concept embeddings) are reliable. For this phase of the project, we focused on exploring ways to evaluate the NER+L performed by three models on the entirety of the LTH dataset and the quality of a small subset of the extracted concepts' embeddings.

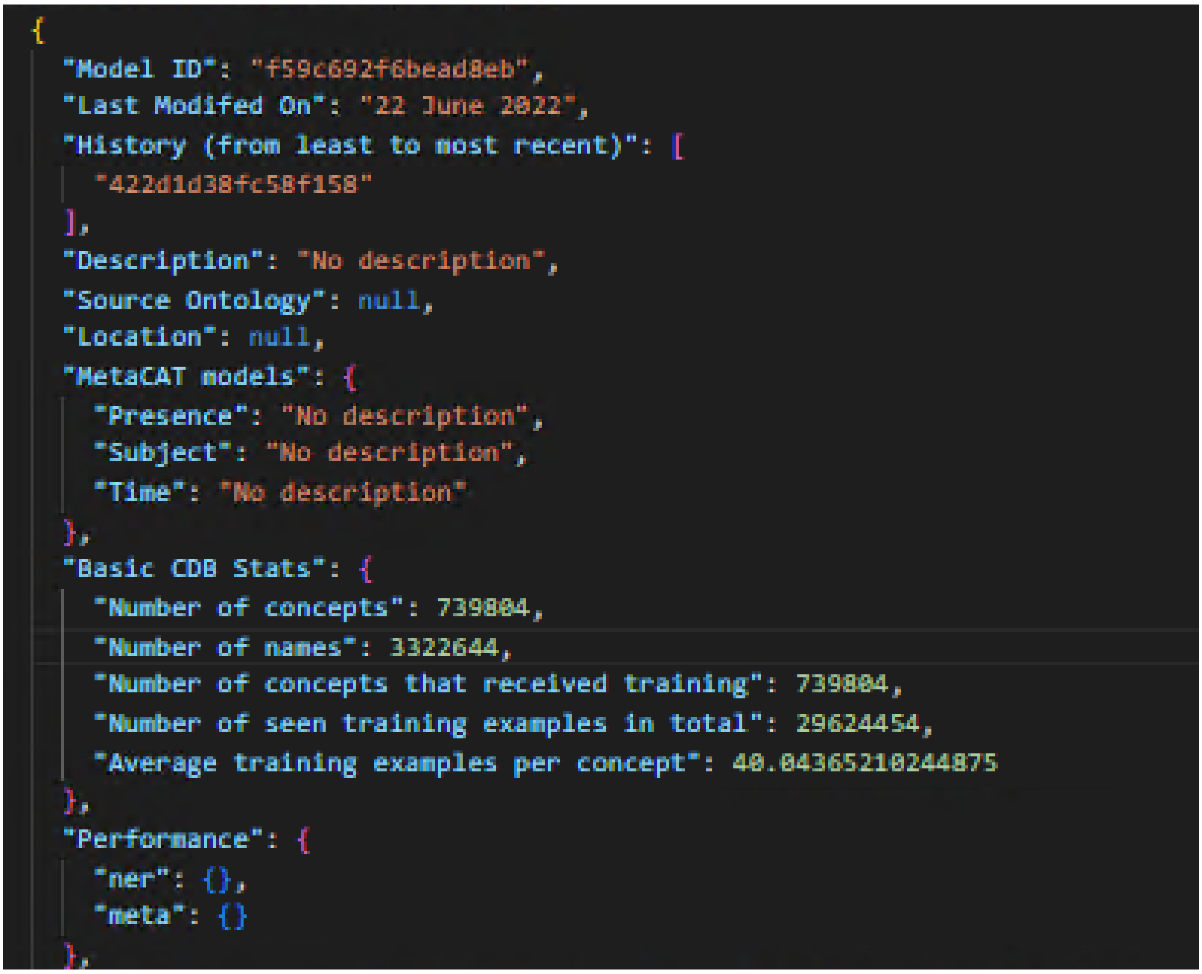
2.2 Corpus

The textual dataset (corpus) used in this project phase is the LTH collection of outpatient clinic letters. This contains 110831 sentences extracted from 3000 outpatient clinic letters from a single consultant neurologist. The dataset has been anonymised, and thus contains no metadata of the original clinic letters, participant gender, age or identifying codes. No pre-annotations of named entities or contexts have been done.

2.3 Models

Each MedCAT text-processing model was loaded as a pack of three elements: 1) the NER+L model (CDB and vocabulary which contain the learned concept embeddings); 2) pre-trained MetaCAT contextualisation (meta-annotation) model; and 3) SciSpacy pre-processing model. NER+L models were produced from (i) unsupervised, (ii) unsupervised + supervised- (fine-tuned) or (iii) unsupervised + unsupervised training by the home institution. They consist of the 300-dimensional embeddings of each concept (concept vectors) that have been learned from a training corpus. A concept vector represents the context (surrounding words) of all its linked named entities. Meta-annotation models contain the parameters of a pre-trained bilateral Long Short-Term Memory Neural Network used to predict the specified context of named entities e.g. negation, temporality and experiencer etc. Pre-processing models contain the parameters for cleaning input text prior to the first-pass NER+L.

Three models were used for NER+L: 1) Public (Section 2.3.1), 2) King's 1.2 (section 2.3.2) and 3) King's 1.4 (section 2.3.3). Refer to Table 2.1 for a summary of the model description. Prior to annotation, the model card (`model_card.json`) (See Figure 2.1) in every model pack and configuration dictionary (`cdb.config`) were inspected.



```
{
  "Model ID": "f59c692f6bead8eb",
  "Last Modifed On": "22 June 2022",
  "History (from least to most recent)": [
    "422d1d38fc58f158"
  ],
  "Description": "No description",
  "Source Ontology": null,
  "Location": null,
  "MetaCAT models": {
    "Presence": "No description",
    "Subject": "No description",
    "Time": "No description"
  },
  "Basic CDB Stats": {
    "Number of concepts": 739804,
    "Number of names": 3322644,
    "Number of concepts that received training": 739804,
    "Number of seen training examples in total": 29624454,
    "Average training examples per concept": 40.04365210244875
  },
  "Performance": {
    "ner": {},
    "meta": {}
  },
```

Figure 2.1: Example MedCAT model card.

Model Name & Last Modified Date	**MedCAT Version**	**Source Ontology**	**Training Dataset**	**Training Received**	**Total Concepts**	**Total Names**	**Concepts Trained**	**Total Training Examples Seen**	**Average Training Examples Per Concept**	**Similarity Threshold**	**Concept Vector Types & Window Size**	**Concept Filters**	**Meta-Annotation Task(s)**
Public 16.2.22	1.2.dev361	SNOMED INT UMLS	MIMIC-III	Unsupervised	354448	2049216	29674	20585988	693.74	0.2	Xlong (27) Long (18) Medium (9) Short (3)	None	Status
King's 1.2 22.6.22	1.2.9.dev13	*Not provided*	MIMIC-III KCH	Unsupervised + Supervised	739804	3322644	739804	29624454	40.04	0.3	XXXlong* XXlong* Xlong (27) Long (18) Medium (9) Short (3)	Present	Presence, Time, Subject
King's 1.4 12.10.22	1.3.1.dev27	SNOMED UK Clinical SNOMED Drug UMLS	MIMIC-III KCH	Unsupervised + Supervised	760282	3079903	22542	150049175	6656.43	0.3	Xlong (27) Long (18) Medium (9) Short (3)	None	Presence, Time, Subject

Table 2.1: Summary of three MedCAT NER+L model description. *Context window size not provided

2.3.1 Public SNOMED-CT Model Trained on MIMIC-III Dataset

This publicly-available model (available from https://github.com/CogStack/MedCAT) was pretrained unsupervised by KCH on the MIMIC-III dataset (predominantly consisting of Intensive Care health records), using the SNOMED International ontology enriched with concepts from UMLS[46]. It was released on 22.6.2022. There are 354448 concepts in the concept database. Out of those, only 29674 received training. There are a total of 20585988 seen training examples, with 693.74 training examples per concept. A detected candidate entity is linked to a concept if the similarity between its context vector and a learned concept vector exceeds a threshold of 0.2. Four types of 300-dimension concept vectors were produced i.e. Xlong (context window of 27 tokens), Long (context window of 18), Medium, (context window of nine) and short (context window of three). It was trained on one meta-annotation task i.e. 'Status', whereby concepts were given either a value of 'Affirmed' if detected or 'Other' if negated or hypothetical.

2.3.2 King's Model 1.2.9

Transferred from King's College London, this model was initially trained unsupervised on the MIMIC-III dataset then finetuned on KCH clinical documents, using the SNOMED International ontology enriched with UMLS. There are 739804 concepts in the concept database. All concepts received training. There are a total of 29624454 seen training examples, with 40.04 training examples per concept. The similarity threshold was set to 0.3. Six types of 300-dimensional concept vectors were produced i.e. XXXlong (context window not provided), XXlong (context window not provided), Xlong (context window of 27 tokens), Long (context window of 18), Medium, (context window of nine) and short (context window of three). Meta-annotation was performed on the tasks of Presence (existence of named entity), Time (if entity existed now, in the past or in the future) and Subject (person experiencing entity). The finetuning process and evaluation metrics are not available in the model documentation.

2.3.3 King's Model 1.4

Transferred from King's College London, Model 1.4 was initially trained unsupervised on MIMIC-III dataset then finetuned on KCH clinical documents, using the SNOMED International ontology enriched with UMLS. No concept filters were applied. There are 760282 concepts in the concept database. 22542 concepts received training. There are a total of 150049175 seen training examples, with 6656.43 training examples per concept. The similarity threshold was set to 0.3. Four types of 300-dimension concept vectors were produced i.e. Xlong (context window of 27 tokens), Long (context window of 18), Medium, (context window of nine) and short (context window of three). Model 1.4 was trained on three meta-annotation tasks of Presence, Time and Subject.The finetuning process and evaluation metrics are not available in the model documentation.

2.4 Exploratory Data Analysis

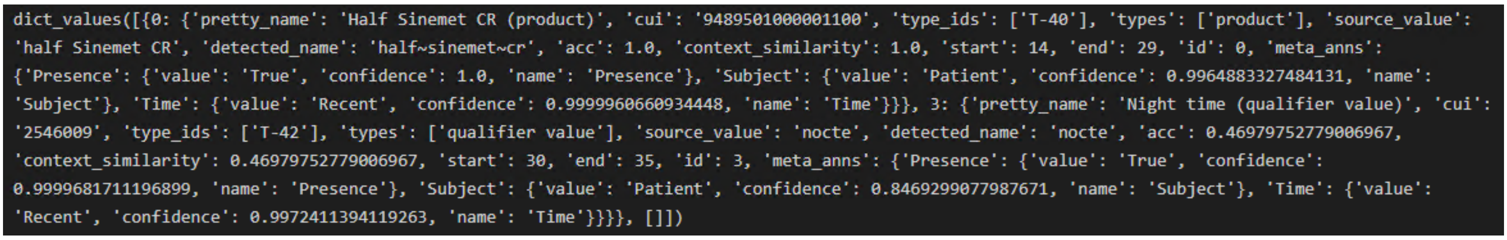
The raw data was loaded into a Python Pandas dataframe and inspected for length, missing values and unique values. Minimum, maximum and average character as well as word lengths were examined. Frequency of documents versus lengths by word and character were plotted.

2.5 Data and annotator preparation

All documents, regardless of length, were included for annotation. The raw data was loaded as a Python Pandas dataframe and inspected. Input data was prepared by populating a list of tuples of two elements: the document ID, followed by the document text from the dataframe. The vocabulary (`vocab.dat`), concept database (`cdb.dat`), and MetaCAT models were loaded from their respective model packs. All three model components were initialised. The main class from MedCAT used for concept annotation, `CAT`, was then initialised with the model pack components loaded earlier.

2.6 NER+L and Meta-Annotation

For all models, NER+L and meta-annotation were performed on input data using the function `cat.multiprocessing()` with a batch size of 50000 characters and number of processors set to eight. The output of applying this function to the corpus is a multilayered nested Python dictionary, with the most superficial key being the document ID. Each document ID contained a Python dictionary of each each named entity detected and linked in the document. Each entity contained a Python dictionary where the keys represent the concept name, concept unique identifier; semantic type identifier; semantic type name; detected name; context similarity; start and end position of detected entity in the document; and a nested dictionary of meta-annotation tasks and their values. An example of the entities for an annotated document, *'He also takes half Sinemet CR nocte'* is given in Figure 2.2.



```
dict_values([{0: {'pretty_name': 'Half Sinemet CR (product)', 'cui': '9489501000001100', 'type_ids': ['T-40'], 'types': ['product'], 'source_value':
'half Sinemet CR', 'detected_name': 'half~sinemet~cr', 'acc': 1.0, 'context_similarity': 1.0, 'start': 14, 'end': 29, 'id': 0, 'meta_anns':
{'Presence': {'value': 'True', 'confidence': 1.0, 'name': 'Presence'}, 'Subject': {'value': 'Patient', 'confidence': 0.9964883327484131, 'name':
'Subject'}, 'Time': {'value': 'Recent', 'confidence': 0.9999960660934448, 'name': 'Time'}}}, 3: {'pretty_name': 'Night time (qualifier value)', 'cui':
'2546009', 'type_ids': ['T-42'], 'types': ['qualifier value'], 'source_value': 'nocte', 'detected_name': 'nocte', 'acc': 0.4697975277906967,
'context_similarity': 0.4697975277906967, 'start': 30, 'end': 35, 'id': 3, 'meta_anns': {'Presence': {'value': 'True', 'confidence':
0.9999681711196899, 'name': 'Presence'}, 'Subject': {'value': 'Patient', 'confidence': 0.8469299077987671, 'name': 'Subject'}, 'Time': {'value':
'Recent', 'confidence': 0.9972411394119263, 'name': 'Time'}}}}, []])
```

Figure 2.2: Example Python nested dictionary of named entities detected and linked with meta-annotation values for document *'He also takes half Sinemet CR nocte'*.

2.7 Processing of Structured Annotation Corpus

To ensure that the `cat.multiprocessing()` function was performed correctly, one randomly-selected input document was compared with the corresponding output. Then, the mapping of each entity to semantic types was checked i.e. one-to-one or one-to-many. If all concepts were mapped in a one-to-one fashion, the structured annotation corpus was rearranged in a flat Pandas dataframe along with the original document for easier inspection. Each column was assessed for length, missing values and unique values. Only 30 most frequent concepts of the type 'Disorder' was selected from this dataframe for further analysis.

2.8 Evaluation

2.8.1 NER+L Coverage

NER+L coverage for each model was assessed by the percentage of documents annotated, total number of entities extracted, number of concepts identified, total number of semantic types and number of unique names of linked entities (see 3.1 for results). Top 30 concepts of type 'Disorder' with most documents mentioning them were plotted.

2.8.2 Learning of Trained Concept Embeddings

Section redacted as manuscript pre-print in progress

2.8.3 3D Visualisation of First Three Principal Components of Long Embeddings

Section redacted as manuscript pre-print in progress

2.8.4 Determining Input Data for Clustering with Principle Component Analysis

Section redacted as manuscript pre-print in progress

2.9 Clustering

Section redacted as manuscript pre-print in progress

2.9.1 K-means Clustering

Section redacted as manuscript pre-print in progress

2.9.2 Affinity Propagation

Section redacted as manuscript pre-print in progress

2.9.3 Agglomerative Clustering with Ward Linkage

Section redacted as manuscript pre-print in progress

2.9.4 Cluster Evaluation

Quantitative

Section redacted as manuscript pre-print in progress

Intrinsic **Section redacted as manuscript pre-print in progress**

Extrinsic **Section redacted as manuscript pre-print in progress**

Qualitative

Section redacted as manuscript pre-print in progress

Chapter 3

Results

3.1 Exploratory Data Analysis

There were 110831 unique sentences in the LTH dataset. Each sentence has a unique document identifier. The mean document length was 15.46 words (SD = 9.79) and ranges from one to 115 words (see Figure 3.1). The average number of characters was 93.29 (SD = 57.34), ranging from one to 676 (see Figure 3.2).

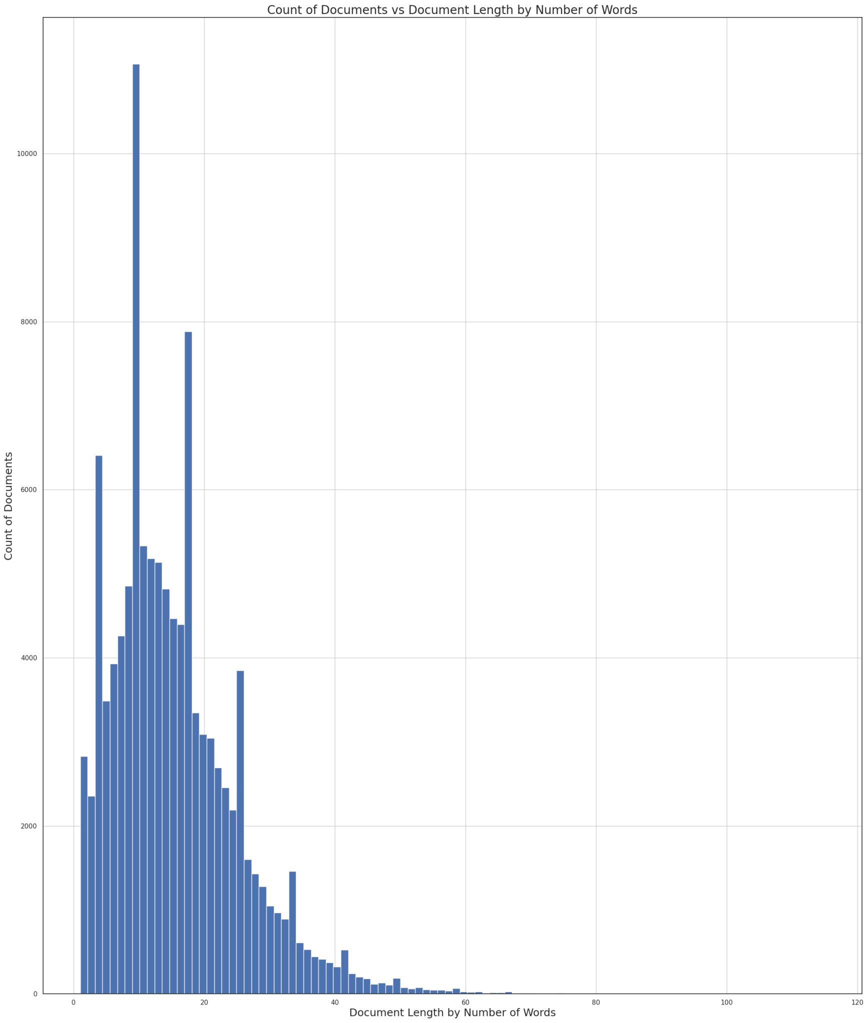


Figure 3.1: Count of Documents vs Document Length by Number of Words

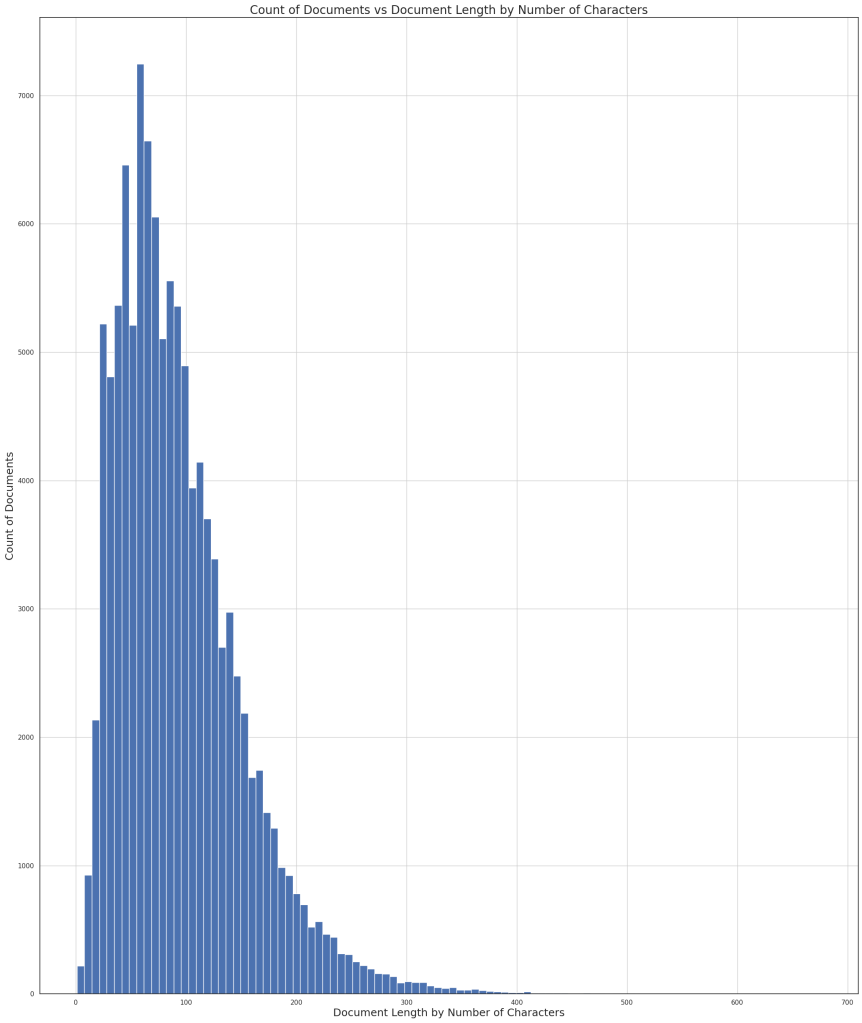


Figure 3.2: Count of Documents vs Document Length by Number of Characters

3.2 NER+L Coverage

NER+L was performed on the LTH corpus using all three models. A summary of the annotation results is given in Table 3.1. All three MedCAT models extracted at least 400000 named entities in over 94% of the clinic letters, with the King's 1.4 model extracting the most (425506) and King's 1.2 extracting the least (402053). Entities were linked to over 9000 concepts for each model. Interestingly, although the King's 1.2 model extracted the least number of entities, these are linked to the highest number of concepts (10117).

Model	Documents Annotated (%)	Entities Extracted	Concepts	Semantic Types	Unique Names
Public	104350 (94.2)	421740	9478	47	9395
King's 1.2	104642 (94.4)	402053	10117	49	10111
King's 1.4	105355 (95.1)	425506	9784	48	9691

Table 3.1: Summary of NER+L results of LTH corpus by three models: 1) Public, 2) King's 1.2 and 3) King's 1.4

3.2.1 Public Model

Using the Public model, 421734 terms were extracted and linked to 9478 concepts which fall under 47 semantic types. The top 30 concepts of the semantic type 'Disorder' is visualised in Figure 3.3. All concepts map one-to-one with the semantic tags. Of the top 30 Disorders, two concept embeddings did not receive training (**Myasthenia gravis** and **Disease caused by severe acute respiratory syndrome coronavirus 2**) and these were discarded from further analysis.

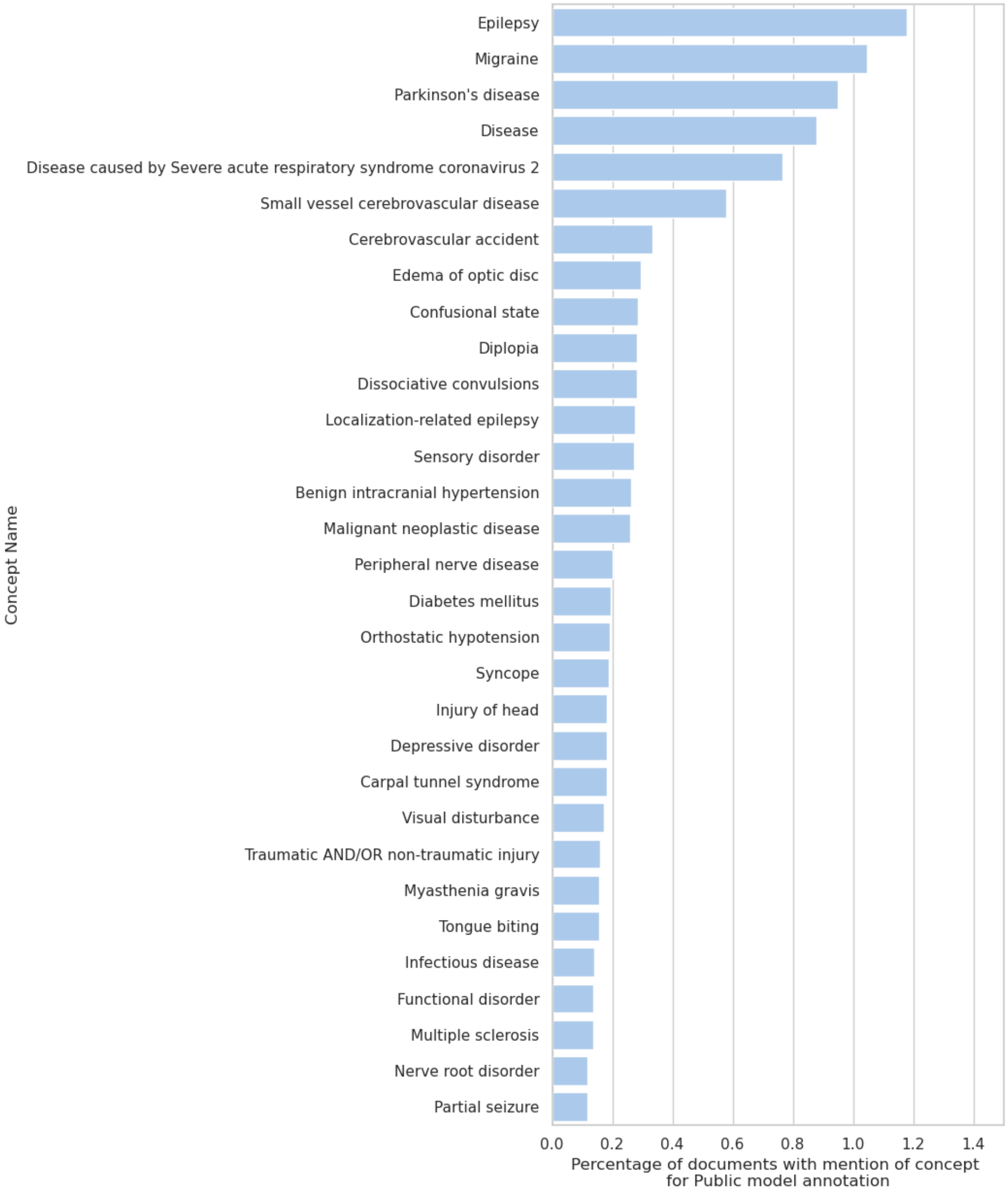


Figure 3.3: Annotation frequency of Public Model for top 30 concepts fo the semantic type 'Disorder'

3.2.2 King's Model 1.2

Following annotation under default configurations, 402053 terms were extracted and linked to 10117 concepts which fall under 49 semantic types. All top 30 concepts of the semantic type 'Disorder' were trained (see Figure 3.4). All concepts map to semantic types in a one-to-one manner. 76 concepts (seven unique identifiers) do not have semantic types and this has been marked as 'Nil' in the structured annotation corpus.

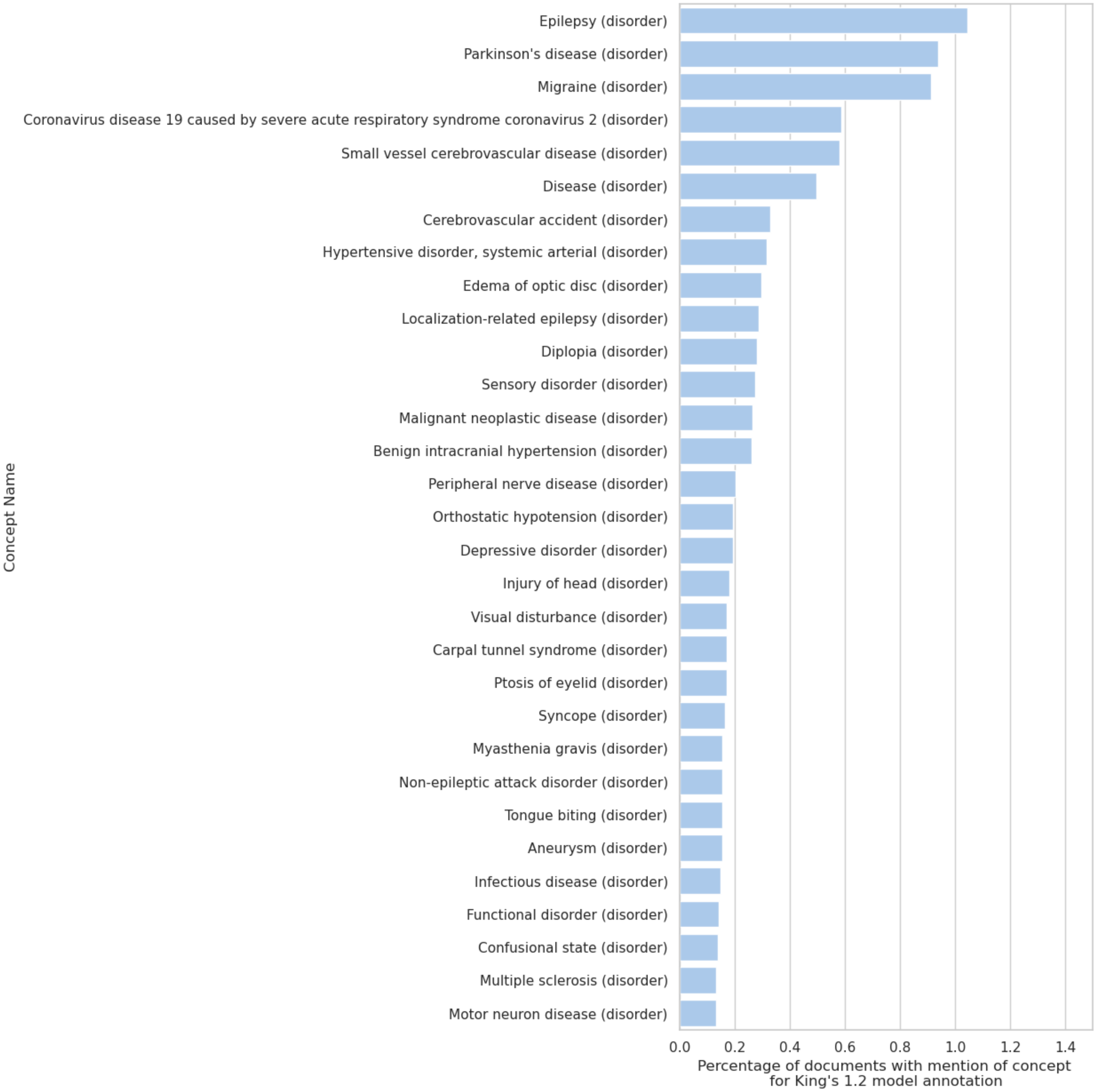


Figure 3.4: Annotation frequency of King's 1.2 model for top 30 concepts for the semantic type 'Disorder'

3.2.3 King's Model 1.4

425506 terms were extracted and linked to 9784 concepts under 48 semantic types. The annotation frequency for the top 30 concepts is shown in Figure 3.5. This model displayed highly rare diseases in the top 30 disorders that we did not expect to appear. On inspection of the underlying text, it appears that the underlying detected values were erroneous i.e. **Congenital malformation** (detected from term 'abnormality'), **Nance-Horan Syndrome** (detected from acronym ‘NHS’), **Lupus erythematosus tumidus** (detected from word ‘let’), **Arsenical keratosis** (detected from word ‘ask’) and **Vascular ectasia of gastric antrum** (detected from word ‘gave’). These erroneous concepts were removed. Out of the 25 remaining concepts, two concept embeddings did not receive training (**Small vessel cerebrovascular disease** and **Sensory disorder**) and these were removed from further analysis.

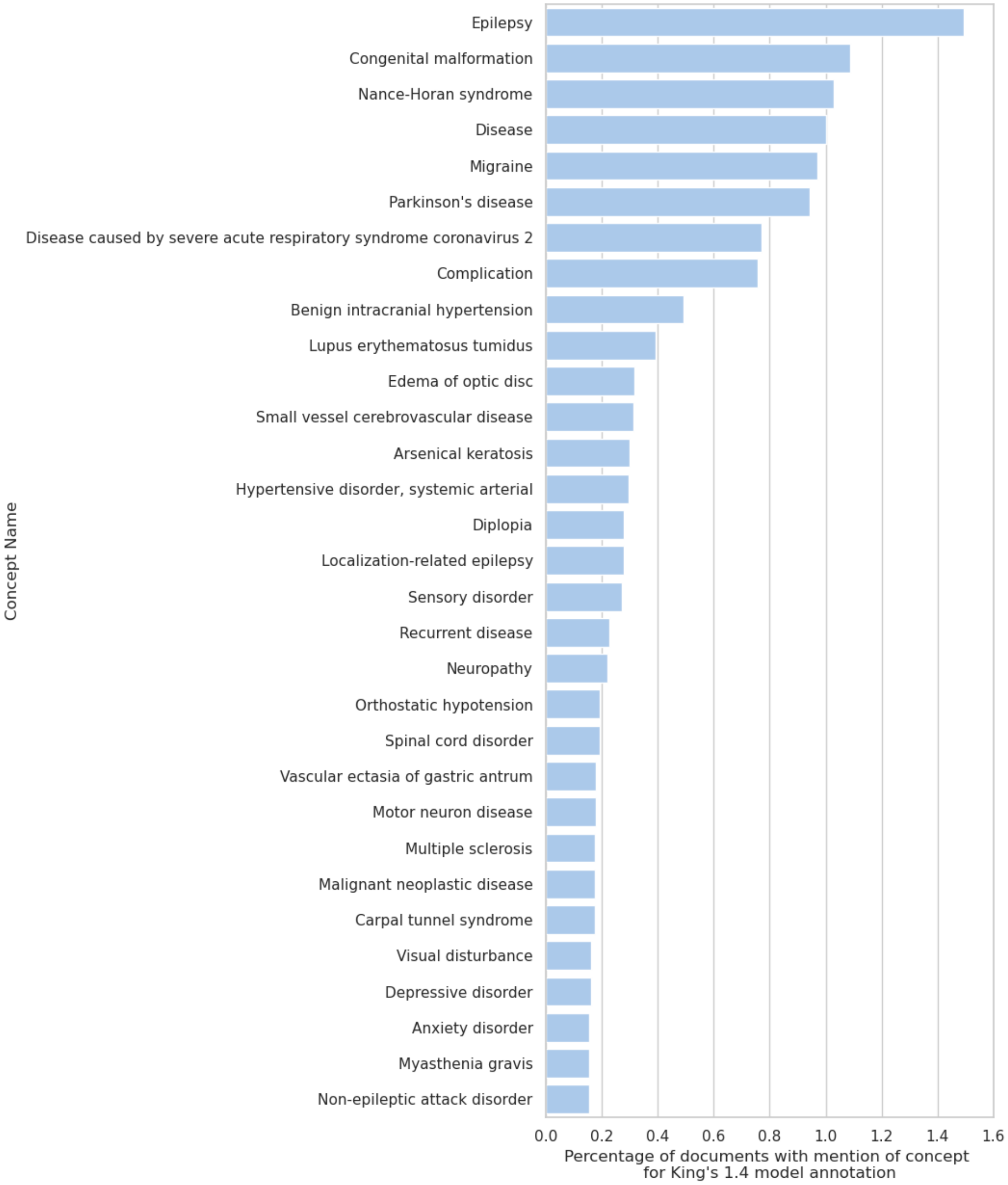


Figure 3.5: Annotation frequency of King’s 1.4 Model for top 30 concepts for the semantic type ‘Disorder’

3.3 Clustering

3.3.1 Quantitative Evaluation

Section redacted as manuscript pre-print in progress

3.3.2 Qualitative Evaluation

Public Model

Section redacted as manuscript pre-print in progress

King's 1.2 Model

Section redacted as manuscript pre-print in progress

King's 1.4 Model

Section redacted as manuscript pre-print in progress

Chapter 4

Discussion

4.1 Discussion

To the best of our knowledge, we are the first organisation in the UK, independent of MedCAT's home institution of KCH, to evaluate pre-trained MedCAT NER+L models i.e. Public SNOMED, King's 1.2 and King's 1.4. These models represent primarily SNOMED-CT concepts (enriched with terms from other ontologies) and their embeddings, learned from the contexts of linked named entities in MIMIC-III and KCH's training data. We first evaluated the coverage of NER+L on LTH Neurology outpatient clinic letters by assessing annotation frequencies of all three models. Then, we proceeded to assess the quality of the learned contexts using a small subset of concept embeddings in the absence of annotated data. In related works, validation of MedCAT NER+L models has all been done using labelled data[14,15,17–19,22]. In healthcare, annotated datasets are scarce, and require a significant amount of resource to produce. Thus, exploring evaluation methods which do not necessarily require human-annotated labels can be of great benefit in low-resource settings. In addition, quantitative evaluation of MedCAT's concept embeddings has not been reported before.

Through reviewing the annotation frequency results (see Section 3.2), we saw that a high total number of entities extracted may not necessarily translate to a high number of linked concepts or unique entity names. This may be due to entities mapping to their concepts in a many-to-one fashion. In addition, each model has slightly different configurations, in that they all have different numbers of concepts and entity names in the CDB, as well as similarity thresholds (see Table 2.1). The King's 1.2 model has the highest number of concepts that received training, which may explain why it extracted the highest number of concepts. Moreover, the CDB for each model is different, and this has a direct bearing on the quantity of unique concepts they can detect. Thus, there is not necessarily a 'good' or 'bad' quantity of annotation. It depends on the MedCAT user as to which concepts they prefer extracting, and to what degree. Given such a target, evaluation of NER+L frequencies can be made more informative.

Upon reviewing the top 30 concepts of type 'Disorder' for every model, we see that many concepts are not necessarily clinically-defined diagnoses which we would expect. Some are highly non-specific e.g. **Disease**, **Sensory disorder**, **Malignant neoplastic disease**, **Confusional state**, **Complication**, and **Infectious disease** and some are simply clinical findings e.g. **Edema of optic disc**, **Diplopia**, **Visual disturbance**, **Tongue biting** and **Ptosis of eyelid**. Conversely, we also found more granular concepts such as **Localisation-related epilepsy** and **Partial seizure** being extracted along with their parent concept i.e. **Epilepsy**. This may be an attribute of the underlying source ontology. In which case, it is important to be aware of the definition of a given semantic type in an ontology as this may not be medically equivalent or relevant. Understanding this will aid in the curation and enrichment of concepts in the CDB. **Section**

redacted as manuscript pre-print in progress

We found that the concept **Functional disorder** was erroneously linked to the entity 'dysfunction' by MedCAT's NER+L method. Clinically, this concept is used to describe conditions presenting with a wide range of symptoms with no underlying anatomical or physiological aberrations, therefore it cannot be linked to the entity 'dysfunction'. The text annotated with this concept also did not show any functional conditions. Inspection of **Functional disorder**'s synonyms in the SNOMED CT browser only revealed the clinically-consistent term 'Functional disturbance'[59]. These findings highly suggest that the CDB was built with incorrect synonyms for **Functional disorder**.

Section redacted as manuscript pre-print in progress

4.2 Limitations and Future Work

There are several limitations in our evaluation of pre-trained NER+L MedCAT models. First, our corpus consists of sentences extracted from clinic letters (with an average length of 16 words) whilst the MedCAT models used were initially trained on clinical notes. Disambiguation during NER+L for our LTH data may have been affected as the named entities in our corpus would have more limited contexts compared to the models' training dataset. This may have affected context similarity values. Annotation frequencies and ranking of concepts could have varied based on document length. Future work may involve exploring the most optimal context length for NER+L (either by varying the context window or the document length itself).

Section redacted as manuscript pre-print in progress

Seventh, we did not assess the frequencies of concepts in view of their meta-annotation values. Any mention of a concept, regardless of its presence, experiencer or temporality is included in the count. Annotation frequencies performed in the future may be done with the meta-annotation values included. Eighth, any linked named entity, regardless of its meta-annotation value, is included by MedCAT in the update of the concept embedding. This may have an effect on annotation frequencies and learned embeddings. Unfortunately, the proportion of false-positive linked entities in the original MedCAT training dataset is not available. Future work can be done to explore the training dataset better. This will enable greater interpretability of NER+L results. Finally, we did not train MedCAT models on LTH data. Therefore, the learned embeddings have not been updated to reflect the context of our dataset. This would have a direct bearing on annotation results. Further training on a target dataset in the future would be ideal to assess if NER+L results improve with further training, and enable evaluation of the rest of MedCAT's functionalities.

Chapter 5

Conclusions

Owing to the clinically-inconsistent results produced by the three pre-trained MedCAT NER+L models evaluated in this paper, we found that they are not yet ready for use in automated clinical coding. Further model training, higher training examples per concept and source ontology curation are required for the NER+L algorithm to detect preferred clinical terms at the desired frequency. Given the inconsistent formatting, documentation and output of the MedCAT models, they are not suitable yet to be used in a standard workflow. Usage of MedCAT requires not only understanding of coding, Python and NLP, but also clinical knowledge. We noted that one single metric is not sufficient to evaluate MedCAT models, and triangulation of results can be made better when employing multiple approaches e.g. intrinsic, extrinsic, qualitative and quantitative. Successful use and evaluation of a biomedical NER+L toolkit such as MedCAT requires strong collaborations between clinicians and data scientists.

References

1. Honeyford, K. *et al.* Challenges and recommendations for high quality research using electronic health records. *Frontiers in Digital Health* **4,** 1–9. ISSN: 2673253X (2022).
2. Crema, C., Attardi, G., Sartiano, D. & Redolfi, A. Natural language processing in clinical neuroscience and psychiatry: A review. *Frontiers in Psychiatry* **13.** ISSN: 16640640 (2022).
3. Locke, S. *et al.* Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care* **38,** 4–9. ISSN: 22108467. https://doi.org/10.1016/j.tacc.2021.02.007 (2021).
4. Dong, H. *et al.* Automated clinical coding: what, why, and where we are? *npj Digital Medicine* **5,** 1–8. ISSN: 23986352 (2022).
5. SNOMED. *5-Step briefing* 2023. https://www.snomed.org/five-step-briefing.
6. Biggin, F. *et al.* Outpatient neurology diagnostic coding : a proposed scheme for standardised implementation, 1–7 (2023).
7. Campbell, S. & Giadresco, K. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Health Information Management Journal* **49,** 5–18. ISSN: 18333575 (2020).
8. Venkatesh, K. P., Raza, M. M. & Kvedar, J. C. Automating the overburdened clinical coding system: challenges and next steps. *npj Digital Medicine 2023 6:1* **6,** 1–2. ISSN: 2398-6352. https://www.nature.com/articles/s41746-023-00768-0 (Feb. 2023).
9. Kaur, R., Ginige, J. A. & Obst, O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Systems with Applications* **213,** 118997. ISSN: 09574174. https://doi.org/10.1016/j.eswa.2022.118997 (2023).
10. Stausberg, J., Lehmann, N., Kaczmarek, D. & Stein, M. Reliability of diagnoses coding with ICD-10. *International Journal of Medical Informatics* **77,** 50–57. ISSN: 1386-5056 (Jan. 2008).
11. Burton, A. How do we fix the shortage of neurologists? *The Lancet Neurology* **17,** 502–503. ISSN: 14744465. http://dx.doi.org/10.1016/S1474-4422(18)30143-1 (2018).
12. Brader, C. *Health care services for neurological conditions - House of Lords Library* May 2022. https://lordslibrary.parliament.uk/health-care-services-for-neurological-conditions/.
13. Spasic, I. & Nenadic, G. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics* **8.** ISSN: 22919694 (2020).
14. Kraljevic, Z. *et al.* Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine* **117.** ISSN: 18732860 (2021).

15. Shek, A. *et al.* Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. *European journal of neurology* **28,** 4090–4097. ISSN: 1468-1331. https://pubmed.ncbi.nlm.nih.gov/34407269/ (Dec. 2021).

16. Bean, D. M., Kraljevic, Z., Shek, A., Teo, J. & Dobson, R. J. B. Hospital-wide natural language processing summarising the health data of 1 million patients. *PLOS Digital Health* **2** (ed Tariq, A.) e0000218. ISSN: 2767-3170. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000218 (May 2023).

17. Bendayan, R. *et al.* 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) , Article e054414. (2022)* **12.** ISSN: 20446055. https://doi.org/10.1136/bmjopen-2021-054414 (Jan. 2022).

18. Noor, K. *et al.* Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals. *JMIR Med Inform 2022;10(8):e38122 https://medinform.jmir.org/2022/8/e38122* **10,** e38122. ISSN: 22919694. https://medinform.jmir.org/2022/8/e38122 (Aug. 2022).

19. Kunz, S., Zgraggen, C. & Sariyar, M. Mapping SNOMED CT Codes to Semi-Structured Texts via an NLP Pipeline. *Studies in Health Technology and Informatics* **295,** 390–393. ISSN: 18798365 (2022).

20. Van Es, B. *et al.* Negation detection in Dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics* **24.** ISSN: 1471-2105. https://pubmed.ncbi.nlm.nih.gov/36624385/ (Dec. 2023).

21. Ariño, H., Bae, S. K., Chaturvedi, J., Wang, T. & Roberts, A. Identifying encephalopathy in patients admitted to an intensive care unit: Going beyond structured information using natural language processing. *Frontiers in Digital Health* **5,** 1085602. ISSN: 2673253X. /pmc/articles/PMC9899891/%20/pmc/articles/PMC9899891/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9899891/ (Jan. 2023).

22. Kraljevic, Z. *et al.* MedCAT – Medical Concept Annotation Tool, 1–25. http://arxiv.org/abs/1912.10166 (2019).

23. Wardle, M. & Spencer, A. Implementation of SNOMED CT in an online clinical database. *Future Hospital Journal* **4,** 126–130. ISSN: 2055-3323 (2017).

24. Nguyen, H. & Patrick, J. Text mining in clinical domain: Dealing with noise. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-Augu,** 549–558 (2016).

25. Wiegreffe, S., Choi, E., Yan, S., Sun, J. & Eisenstein, J. Clinical concept extraction for document-level coding. *BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task,* 261–272 (2019).

26. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology* **2,** 230–243. ISSN: 20598696 (2017).

27. Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A. & Hersh, W. R. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association* **17,** 646–651. ISSN: 10675027 (2010).

28. Pakhomov, S. V., Buntrock, J. D. & Chute, C. G. Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. *Journal of the American Medical Informatics Association* **13,** 516–525. ISSN: 10675027. https://academic.oup.com/jamia/article/13/5/516/734238 (2006).

29. Zeng, Z., Deng, Y., Li, X., Naumann, T. & Luo, Y. Natural Language Processing for EHR-Based Computational Phenotyping HHS Public Access. *IEEE/ACM Trans Comput Biol Bioinform* **16,** 139–153 (2019).

30. Sonabend W, A. *et al.* Automated ICD coding via unsupervised knowledge integration (UNITE). *International Journal of Medical Informatics* **139,** 104135. ISSN: 18728243. https://doi.org/10.1016/j.ijmedinf.2020.104135 (2020).

31. Iman, M., Arabnia, H. R. & Rasheed, K. A Review of Deep Transfer Learning and Recent Advancements. *Technologies 2023, Vol. 11, Page 40* **11,** 40. ISSN: 2227-7080. https://www.mdpi.com/2227-7080/11/2/40/htm%20https://www.mdpi.com/2227-7080/11/2/40 (Mar. 2023).

32. Gupta, N. A Pre-Trained Vs Fine-Tuning Methodology in Transfer Learning. *Journal of Physics: Conference Series* **1947,** 012028. ISSN: 1742-6596. https://iopscience.iop.org/article/10.1088/1742-6596/1947/1/012028%20https://iopscience.iop.org/article/10.1088/1742-6596/1947/1/012028/meta (June 2021).

33. Durrani, N., Sajjad, H. & Dalvi, F. How transfer learning impacts linguistic knowledge in deep NLP models? *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 4947–4957. https://arxiv.org/abs/2105.15179v1 (May 2021).

34. Kavuluru, R., Rios, A. & Lu, Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine* **65,** 155–166. ISSN: 18732860. http://dx.doi.org/10.1016/j.artmed.2015.04.007 (2015).

35. Kavuluru, R., Han, S. & Harris, D. Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7884 LNAI,** 77–88. ISSN: 03029743. https://link.springer.com/chapter/10.1007/978-3-642-38457-8_7 (2013).

36. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J. & Krallinger, M. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020. *CEUR Workshop Proceedings* **2696,** 22–25. ISSN: 16130073 (2020).

37. García-Santa, N. & Cetina, K. FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding. *CEUR Workshop Proceedings* **2696,** 22–25. ISSN: 16130073 (2020).

38. Wen, C., Chen, T., Jia, X. & Zhu, J. Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. *Data Intelligence* **3,** 402–417. ISSN: 2641435X (2021).

39. Song, H. J., Jo, B. C., Park, C. Y., Kim, J. D. & Kim, Y. S. Comparison of named entity recognition methodologies in biomedical documents. *BioMedical Engineering Online* **17,** 1–14. ISSN: 1475925X. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0573-6 (Nov. 2018).

40. Sharma, R., Morwal, S. & Agarwal, B. Named entity recognition using neural language model and CRF for Hindi language. *Computer Speech & Language* **74,** 101356. ISSN: 0885-2308 (July 2022).

41. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes* **30,** 3–26. ISSN: 0378-4169. https://www.researchgate.net/publication/44062524_A_Survey_of_Named_Entity_Recognition_and_Classification (Aug. 2007).

42. Lossio-Ventura, J. A., Boussard, S., Morzan, J. & Hernandez-Boussard, T. Clinical named-entity recognition: A short comparison. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019,* 1548–1550 (2019).

43. Goulart, R. R. V., Strube de Lima, V. L. & Xavier, C. C. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society* **17,** 103–116. ISSN: 16784804 (2011).

44. Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: An introduction. *Journal of the American Medical Informatics Association* **18,** 544–551. ISSN: 10675027 (2011).

45. *ICD-10 Version:2019* https://icd.who.int/browse10/2019/en.

46. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* **32.** ISSN: 03051048 (Jan. 2004).

47. Hachey, B., Radford, W., Nothman, J., Honnibal, M. & Curran, J. R. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence* **194,** 130–150. ISSN: 0004-3702 (Jan. 2013).

48. Shen, W., Wang, J. & Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* **27,** 443–460. ISSN: 10414347 (Feb. 2015).

49. Zheng, J. G. *et al.* Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making* **15,** 1–9. ISSN: 14726947. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-15-S1-S4 (May 2015).

50. Landolsi, M. Y., Romdhane, L. B. & Hlaoua, L. Medical Named Entity Recognition using Surrounding Sequences Matching. *Procedia Computer Science* **207,** 674–683. ISSN: 18770509. https://doi.org/10.1016/j.procs.2022.09.122 (2022).

51. Li, Q. *et al.* Improving Entity Linking by Introducing Knowledge Graph Structure Information. *Applied Sciences (Switzerland)* **12,** 2702. ISSN: 20763417. https://www.mdpi.com/2076-3417/12/5/2702/htm%20https://www.mdpi.com/2076-3417/12/5/2702 (Mar. 2022).

52. Abdurxit, M., Tohti, T. & Hamdulla, A. An Efficient Method for Biomedical Entity Linking Based on Inter- and Intra-Entity Attention. *Applied Sciences 2022, Vol. 12, Page 3191* **12,** 3191. ISSN: 2076-3417. https://www.mdpi.com/2076-3417/12/6/3191/htm%20https://www.mdpi.com/2076-3417/12/6/3191 (Mar. 2022).

53. French, E. & McInnes, B. T. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics* **137,** 104252. ISSN: 1532-0464 (Jan. 2023).

54. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.* http://ronan.collobert.com/senna/ (2013).

55. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data 2016 3:1* **3,** 1–9. ISSN: 2052-4463. https://www.nature.com/articles/sdata201635 (May 2016).

56. Neumann, M., King, D., Beltagy, I. & Ammar, W. *ScispaCy: Fast and robust models for biomedical natural language processing* in *BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task* (Association for Computational Linguistics (ACL), 2019), 319–327. ISBN: 9781950737284.

57. *MedCAT Tutorial | Part 3.1 Building a Concept Database and Vocabulary.ipynb - Colaboratory* https://colab.research.google.com/github/CogStack/MedCATtutorials/blob/main/notebooks/introductory/Part_3_1_Building_a_Concept_Database_and_Vocabulary.ipynb#scrollTo=OgMSGHyhk7gv.

58. Chandrabalan, V. & Dobson, S. *LANDER - Lancashire Data Science Environment* 2022. http://northwest-lsc-tre.surge.sh/#/7/4.

59. SNOMED International. *SNOMED CT - Functional disorder (disorder)* 2017. https://termbrowser.nhs.uk/?perspective=full&conceptId1=386585008&edition=uk-edition&release=v20230607&server=https://termbrowser.nhs.uk/sct-browser-api/snomed&langRefset=999001261000000100,999000691000001104.