Master's Thesis

(A)I Will Catch You When You Fall: Fall Registry Development of Hip Fracture Patients using Natural Language Processing

Michelle Chan

(6976794)

MSc Artificial Intelligence

Utrecht University
Department of Information and Computing Sciences
Faculty of Science

Harvard Medical School
Massachusetts General Hospital
Department of Orthopaedic Surgery

Supervised by: Tejaswini Deoskar (Utrecht University)
Soheil Ashkani (Harvard Medical School)
2^{nd} examiner: Pablo Mosteiro Romero (Utrecht University)

November 6, 2023

Abstract

Falls present a pressing public health concern, necessitating a comprehensive grasp of their occurrences and underlying causes. Patient registries are invaluable resources for understanding disease progression and clinical practices, yet their development through conventional methods, such as the ICD framework, might lead to an underestimation of fall frequencies due to coding limitations. Natural language processing (NLP) emerges as a promising solution for automating the analysis of clinical notes, enabling the creation of a comprehensive fall registry.

This thesis addresses the complexities of fall registry development, focusing specifically on hip fracture patients. It encompasses well-established tasks such as identifying fall occurrences, as well as innovative challenges involving the extraction of fall mechanisms and the classification of fall impact. We not only delineated these tasks by developing comprehensive guidelines but also gauged their difficulty through a meticulous comparison of medical expert annotations with layman's annotations. This comparison revealed that tasks related to fall occurrences and fall mechanisms were straightforward and required no medical inference, unlike fall impact classification. The annotations for the initial two tasks served as the foundation for our modelling experiments.

In fall detection, we utilised a combination of rule-based techniques, weakly supervised machine learning, and BERT, all of which outperformed traditional ICD codes by detecting 98% of falls — a significant 78% improvement. Additionally, we delved into fall mechanism extraction using pre-trained QA models, with the best model achieving a F1 score of 0.34. By setting a retrieval threshold above 0.50 F1, we identified 21% more fall mechanisms compared to ICD coding.

These experiments underscore the substantial potential of natural language processing (NLP) in significantly enhancing the fall registry development process to facilitate future research regarding falls.

Acknowledgments

I am immensely grateful to my supervisors, Tejaswini and Soheil, for their expert guidance and support throughout my research journey. They provided me with valuable insights into natural language processing and deepened my understanding of orthopaedics.

I extend my sincere appreciation to Atta and Evan for their unwavering dedication to the annotation study, despite the demanding and time-consuming nature of the task. Their meticulousness and hard work played a pivotal role in shaping the outcomes of this project, and I am truly grateful for their contributions.

I would like to express my thanks to Iris, Tim, and Joppe for their unwavering support and intellectual sparring throughout this journey. Their encouragement helped me to push through even when we were over 5,000 kilometres apart.

Lastly, I am grateful for the opportunity to experience this final part of my student journey in Boston, a city that now holds a special place in my heart.

Table of Contents

Chapter 1

Introduction

Falls are a major concern for public health. They are one of the leading causes of accidental injury and the top cause of death for individuals over the age of 65 (Moreland et al., 2020). Each year, 30% of the elderly in the United States experience falls, resulting in 3 million emergency room visits, 300,000 hospitalisations for hip fractures, and 30,000 fatalities (Bergen et al., 2016; Moreland et al., 2020). The medical expenses associated with fall-related injuries are estimated to be around $50 billion and have financial implications for both patients and the healthcare system. As the population ages, the prevalence of falls and their associated medical costs are expected to increase significantly (Florence et al., 2018). Understanding the nature of falls and fall-related injuries can help reduce financial costs and facilitate the development of effective prevention strategies and treatments.

Patient registries are a powerful tool that can provide valuable insight into disease progression, treatment outcomes, and clinical practice (Schmidt et al., 2015; Vollmer et al., 1999). Defined by Gliklich et al. (2020) as "organised systems that collect uniform data for a population that shares a particular disease, condition, or procedure", these registries exist for a wide range of diseases and conditions, including cancer, diabetes, and blindness (Workman, 2013; Parkin, 2006; Tan et al., 2019). One common approach to developing patient registries involves utilising administrative codes assigned to patient encounters, such as International Classification of Diseases (ICD) diagnosis codes (Parkin, 2006). This method has also been applied in recent studies focused on falls, where patients who have experienced falls are identified using ICD codes to establish related registries, such as trauma or injury registries (Khorgami et al., 2018;

Unguryanu et al., 2020; Sumrein et al., 2017).

Compared to patient registries for diseases such as cancer Gjerstorff (2011); Bilimoria et al. (2008), developing a fall registry using the ICD method may lead to an underestimation of the true frequency of falls (Tremblay et al., 2009). Although the ICD framework includes specific codes for falls, such as "fall due to ice and snow", their usage is limited (McKenzie et al., 2006), and healthcare workers may not be familiar with them. Falls are generally not considered as standalone diseases or conditions. In certain instances, only the resulting injury, such as a hip fracture, is coded rather than the underlying cause. This discrepancy arises because the ICD framework was primarily designed for billing purposes rather than the classification of diseases and conditions, despite its name implying otherwise (Jensen et al., 2012).

When structured data or codes such as ICD codes are limited, as seen in the case of falls, researchers and healthcare workers can shift to reviewing clinical notes. This process, known as chart review, involves analysing free-texts such as discharge summaries, consultations, and physician progress notes to find relevant information (Gliklich et al., 2020). Inspection of the detailed description of a patient's clinical journey, including the reason for admission, prior hospital encounters and lists of prescribed medication, has increased the proportion of adverse events captured Hill et al. (2010); Olsen et al. (2007). However, chart reviews can be time-consuming and difficult due to the lengthiness of the data and the (lack of) structure it may have. Relevant information, which is often embedded within a single sentence (e.g."patient fell from the stairs"), can be hidden between pages of test results and medication lists. To keep up with the growing volume of notes and alleviate the burden on healthcare workers, natural language processing (NLP) can be utilised to automate the reviewing process of clinical notes to develop a fall registry (Gliklich et al., 2020).

Previous studies have successfully developed supervised models that effectively identify fall occurrences in clinical notes (Patterson et al., 2019; Shiner et al., 2020; Fu et al., 2022). However, the development of these models required a time-consuming and labour-intensive annotation process to generate training data (Tohira et al., 2022). Furthermore, the lack of de-identification incorporated in the data pre-processing lim-

ited the re-usability of the annotated data due to privacy regulations, impeding the deployment of new models and limiting their practical implementation in real-world scenarios.

While the current generation of fall models has proven adept at identifying fall occurrences, they grapple with limitations in capturing fine-grained fall-related ICD codes that encompass the specific mechanism behind falls (i.e. the manner in which the patient fell) (Tremblay et al., 2009) and the impact of falls (i.e. the force or energy exerted on the body during a fall). Understanding the contextual factors and implications of falls is crucial for developing effective preventive strategies (Unguryanu et al., 2020). Moreover, it is essential for evaluating a patient's susceptibility to future falls, especially since individuals who have experienced high-impact falls face an elevated risk of recurrent falls and severe injuries (Leucht et al., 2009). Although Tremblay et al. (2009) has recognised the potential for extracting the mechanism of falling as an area for future research, the automatic extraction of these fall mechanisms and the classification of fall impact from clinical notes remain uncharted territories within the current literature.

The primary objectives of this thesis are twofold. Firstly, it aims to establish a robust foundation for advancing comprehensive NLP models for registry development regarding falls. This foundation is constructed through the development of annotation guidelines, which serve the dual purpose of ensuring consistency in the base task of identifying fall occurrences and laying the groundwork for two novel tasks: identifying fall mechanisms and assessing fall impact. This is achieved through an annotation study designed to provide insights into the difficulty of these tasks by comparing the annotations of a layman with those of medical experts.

The second objective of this thesis is to develop a comprehensive fall registry for hip fracture patients, one of the most relevant study populations in the field of orthopaedics due to the high incidence of hip fractures resulting from falls (Parkkari et al., 1999). To identify the most effective approach for developing this registry, various models were developed using the aforementioned tasks. These models encompass rule-based, machine learning, and BERT methods for identifying fall incidents, as well as pre-

trained question-answering models for extracting fall mechanisms. Given the novelty and difficulty of the fall impact task, this thesis exclusively focuses on identifying fall occurrences and extracting fall mechanisms from clinical notes to construct the registry.

1.1 Thesis Outline

This thesis is organised as follows. Chapter 2 describes the related work regarding fall-related NLP research, including prior work on identifying falls in clinical notes. Chapter 3 describes the data that was used for this thesis, how it was collected, and eventually cleaned and pre-processed. Chapter 4 outlines the annotation study that was conducted to assess the difficulty of the fall tasks of identifying fall occurrence, extracting fall mechanism, and classifying fall impact, to develop a guideline for fall annotations, and utilise the annotations for developing the evaluation dataset. Chapter 5 describes the methodology for the task of identifying fall occurrence and discusses the findings and limitations. Chapter 6 describes the methodology for the task of fall mechanism extraction and also discusses the findings and limitations. Chapter 7 concludes this thesis, and addresses the objectives and the contributions of this thesis.

Chapter 2

Related Work

In recent years, there has been a growing interest in NLP for extracting relevant information from electronic health records (EHRs) to develop clinical patient registries (Palmer et al., 2019; Savova et al., 2008).

Section 2.1 will provide a comprehensive background on EHRs, the main challenges in working with them, and the pre-processing steps undertaken in prior research.

The focus of fall registry development has mostly been on identifying fall occurrences in EHRs, as discussed in section 2.2. However, current algorithms are unable to extract detailed information regarding the manner in which a fall occurred, as pointed out by Tremblay et al. (2009). To address this issue, we will explore the relevance of event extraction to fall analysis, as explained in section 2.3.

2.1 Electronic Health Records (EHRs)

EHRs have emerged as vital tools in modern healthcare, serving as digital repositories for patients' medical histories encompassing diagnoses, treatments, laboratory tests, and medication lists (Dalianis, 2018). These records contain a vast amount of valuable clinical data, and their increasing accessibility for secondary purposes, such as clinical research, has made it imperative to develop methodologies that can protect patient privacy while enabling collaborative research (Meystre et al., 2014).

In contrast to open-source datasets, EHRs often contain Protected Health Information (PHI)[1] with personal identifiers such as names, phone numbers, and social security numbers (Norgeot et al., 2020). Access to PHI is limited by the Health Insur-

[1] Appendix A

ance Portability and Accountability Act (HIPAA), which results in limited availability of open-source clinical annotated data, standardised pre-processing procedures, and the reuse of datasets and models (Pomares-Quimbaya et al., 2019; Uzuner et al., 2006). To overcome this, automatic de-identification of medical documents has become crucial for the development of machine learning models. Subsection 2.1.1 will provide an overview of text de-identification methodologies in the clinical domain.

In addition to the sensitive content contained within EHRs, another notable characteristic is the format in which EHR data is typically recorded (Uzuner et al., 2007). Existing literature recognises two distinct types of data in EHRs: structured and unstructured (Dalianis, 2018). Structured data exist in relational databases and include patient demographics, administrative billing codes (e.g. ICD and CPT codes), medication lists, and laboratory tests. On the other hand, unstructured data consist of free-text narratives such as discharge summaries, consultations, and physician progress notes (Sarwar et al., 2022; Dalianis, 2018). In practice, the majority of information is contained within the free-texts (Griffon et al., 2014), including data that are theoretically defined as structured (Gobbel et al., 2022; Skentzos et al., 2011).

Clinical narratives, despite being in free-text format, often follow a conceptual or electronic template that organises the text into general sections (Tepper et al., 2012). When retrieved from EHR databases, these templates are mapped into plain texts resulting in examples such as "Patient ID: 12345" and "HISTORY OF PATIENT ILLNESS Patient fell off the stairs". Although these examples demonstrate the presence of structured sections, previous studies have only focused on identifying falls at the document- or sentence-level (Patterson et al., 2019; Shiner et al., 2020; Fu et al., 2022). However, there is potential in structuring free texts into sections to enhance performance in clinical extraction tasks, such as clinical registry development (Edinger et al., 2017). In subsection 2.1.2, we will provide an overview of this approach.

2.1.1 De-identifying Clinical Notes

De-identification of clinical notes, involving the removal or substitution of PHI, has been an ongoing task since the late 1990s (Sweeney, 1996).

Early approaches relied on rule-based approaches to identify PHI (Hartman et al., 2020; Beckwith et al., 2006). These methods required minimal annotated training data and targeted specific entities. However, rule-based models faced challenges such as lexical ambiguity between PHI and non-PHI entities, the presence of out-of-vocabulary PHI, and limitations in generalisability and scalability (Uzuner et al., 2007).

In contrast, machine learning-based methods, such as conditional random field (CRF) and support vector machine (SVM), can learn to identify PHI patterns from annotated instances and offer improved generalisability and scalability (Liu et al., 2017). While hybrid approaches that combine rule-based and machine learning techniques have initially achieved the best results (Deleger et al., 2013), recent advancements have shown that recurrent neural networks (Dernoncourt et al., 2017) and transformers (Chambon et al., 2022) can achieve state-of-the-art performances on various open-source benchmark datasets, such as MIMIC (Johnson et al., 2023) and i2b2 (Uzuner et al., 2006).

Various methods have been proposed for de-identifying PHI after the identification stage, including removal (Neamatullah et al., 2008), substitution with a higher-level category (Sweeney, 1996) or synthetic PHI (Chambon et al., 2022), and swapping them between texts (Dalenius & Reiss, 1982; Douglass et al., 2004).

Recent fall-related NLP research has not incorporated text de-identification (Fu et al., 2022; Tohira et al., 2022), which may be caused by the limited availability of open-source de-identification models or the concerns about potential information loss (Li & Qin, 2017). However, it has been suggested that the loss can be minimised, and the integration of text de-identification can help in advancing healthcare by facilitating the development and reuse of models (Meystre et al., 2014), which is currently constrained by a chicken and egg problem: effective system development requires access to clinical records, but making clinical records available for research (even for de-identification) requires them to be de-identified first (Uzuner et al., 2007).

In an endeavour to solve this issue and to set standard practice, this thesis takes a first step by incorporating a de-identification process in fall-related NLP research, using a state-of-the-art open-source transformer developed by Chambon et al. (2022) which has also not yet been applied to research beyond benchmarks for de-identification

tasks.

2.1.2 Structuring Free-Texts into Sections

Section segmentation, also known as section identification, is the process of identifying and marking the boundaries of consecutive sentences that share a common theme or topic (Pomares-Quimbaya et al., 2019). This process is crucial in unveiling underlying structures of texts, thereby enhancing information extraction tasks (Tepper et al., 2012). However, section segmentation in clinical texts poses significant challenges due to the diverse types of clinical notes, variations in software usage among different hospitals, and various naming conventions potentially caused by the individual clinicians' modification flexibility.

Clinical notes typically adhere to a free-text template from EHR systems, wherein sections can be identified by explicit section indicators such as headings. Headings usually follow a specific format, such as a first capitalised letter followed by lowercase letters and a colon (e.g. "Conclusion:"), all capitalised (e.g. "CONCLUSION:"), or all capitalised without the colon. However, in some cases, clinicians may omit section headings or replace them with paragraph breaks (Cho et al., 2003), indicating implicit sections. This often occurs when clinicians use a writing framework (Mowery et al., 2012) or write a document from scratch.

Previous work in this field has focused on developing systems that identify both implicit and explicit sections through a combination of heuristics and machine learning. Cho et al. (2003) employed a hybrid method that incorporated heuristics and machine learning. They collected candidate sections based on conventions indicating section headings, such as the presence of a colon or the use of all capitalised letters. These candidates were manually divided into predefined classes, including report header, procedure, history, and conclusion. By utilising lexical patterns and statistics on the mean length of headings, Cho et al. (2003) trained a classifier that achieved accuracy ranging from 0.92 to 0.99 on identifying implicit and explicit sections. Denny et al. (2009) identified candidate headings in History & Physical examination (H&P) notes using lexical pattern matching. They asked clinicians to create a list of twenty-nine section head-

ings and calculated Bayesian probabilities for each sentence to determine its association with a given section, eliminating the need for manual annotation and achieving notable precision and recall scores above 0.90. Li et al. (2010a) utilised a Hidden Markov Model (HMM) to classify sections in medical records by finding an optimal sequence of section categories. They mapped section headers to 15 manually selected general section categories and achieved an accuracy of 0.70. Tepper et al. (2012) utilised a statistical segmentation approach based on formal rules to address the reliance on handcrafted rules for boundary detection using a two-step approach: first, classifying lines using the "BIO" token tagging scheme (Gu et al., 2021), including features which determine whether a word is all capitalised; and second, machine learning-based section labelling using heading and body features, such as average length. This method demonstrated good performance for both implicit and explicit sections in discharge summaries.

While previous approaches have shown promising results in identifying and marking implicit and explicit sections, the availability of section parsers remains limited. Additionally, considering our data comes from different institutions, we have developed a custom parser for this thesis. Our parser integrates formal rules, drawing inspiration from the lexical patterns used by Cho et al. (2003), along with a line-based two-step approach similar to Tepper et al. (2012).

2.2 Fall Occurrence Detection

The identification of fall incidents in clinical texts has been extensively studied, along with other cohort retrieval studies Savova et al. (2008).

Previous research includes rule-based approaches in which variations of the term "fall" were utilised to identify falls in clinical notes (Patterson et al., 2019; Zhu et al., 2017; Tremblay et al., 2009). Fu et al. (2022) utilised the patterns provided by domain experts or existing studies, whereas Shiner et al. (2020) leveraged the pre-built database Unified Medical Language System (UMLS) to find additional fall-related concepts (Bodenreider, 2004). While rule-based methods have been shown to achieve high performances, they were also susceptible to false positives because of homonyms (e.g.

"his cast fell apart"), negations (e.g. "negative for falling"), and references to fall history and fall risk (e.g. "fall history: YES"). Several studies have used supervised machine learning-based methods to identify fall incidents in clinical notes. McCart et al. (2013) and Tohira et al. (2022) trained classical machine learning models such as logistic regression, SVM, and random forest classifiers, using the text frequency-inverse document frequency (tf-idf) transformation of clinical free texts that were annotated by domain experts. However, Dos Santos et al. (2019) revealed that their Long Short-Term Memory (LSTM) network, which utilised various word embeddings such as *Word2Vec* (Mikolov et al., 2013) to represent the free texts, outperformed classical machine learning classifiers, including random forest and SVM. Fu et al. (2022) demonstrated that transfer learning methods, such as fine-tuning Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), outperformed LSTM and Convolutional Neural Network (CNN) using word embeddings, on both sentence and document level. While these studies achieved high F1 scores, they required a time-consuming and labour-intensive data annotation process (Tohira et al., 2022). This annotation bottleneck is particularly apparent in the clinical domain, where clinical expertise is necessary and the availability of annotated training and benchmarking data is limited due to patient privacy concerns (Shiner et al., 2020; Tremblay et al., 2009; Chapman et al., 2011).

Weak supervision has gained popularity as a means to address the limitations posed by the requirement for annotated data (Hedderich et al., 2020). This technique facilitates researchers to generate data with embedded domain knowledge, commonly in the form of rule-based heuristics or external knowledge bases of which the latter is often referred to as distant supervision (Zhou, 2018). Topaz et al. (2019) utilised a combination of weak supervision and active learning to reduce the time spent on annotations (Elkan & Noto, 2008). In their study, an initial lexicon regarding patient falls was constructed by presenting terms with high semantic similarity to the human input query, allowing a domain expert to select relevant terms in an iterative process. Clinical notes were *weakly* labelled using the lexicon, similar to the rule-based approach, and fed into various machine learning classifiers, including SVMs and decision trees, to classify fall

risk and fall history in a set of unlabelled documents. Positive classifications were manually reviewed and added to the training set to train another round of classifiers. This process was iterated until satisfactory results were achieved. Weakly supervised classifiers outperformed traditional rule-based approaches on F1 metrics and represented themselves as a promising approach to bypass the annotation barrier.

In this thesis, a weak supervision approach is used to annotate the data for fall registry development, following a rule-based method proposed by Chan et al. (2022). This approach leverages WordNet, a lexical database that groups linguistic components into sets of cognitive synonyms (synsets) and interlinks them through conceptual-semantic and lexical relations (Fellbaum, 1998). WordNet can obtain a broader range of related concepts to falls compared to UMLS, which only covers medical concepts. To compare rule-based methods from previous studies to the WordNet approach, and weakly supervised machine learning classifiers similar to the approach used by Topaz et al. (2019); Tohira et al. (2022); McCart et al. (2013), and BERT following Fu et al. (2022), we use them for modelling.

2.3 Clinical Information Extraction

Information Extraction (IE) is a complex and extensively studied area that aims to automatically extract and encode information from text (Wang et al., 2018). In the clinical domain, this involves the extraction of medical concepts, entities, events, their relations, and associated attributes (Wang et al., 2018). However, the automatic extraction of clinical events, such as procedures and diseases, has received relatively limited exploration due to data scarcity (Ma et al., 2023).

Previous studies can be categorised into various approaches. The first approach is framing IE as a Named Entity Recognition (NER) problem, focusing on extracting medical concepts and identifying medical events. Examples of such systems include MedLee (Friedman et al., 1994), cTAKES (Savova et al., 2008), and MedTagger (Liu et al., 2013). However, many of these systems require annotated resources for development, which is an ongoing issue in the clinical domain.

The second approach is Event Extraction (EE), which aims to identify event types, triggers, and associated roles (Doddington et al., 2004). This line of research encompasses both rule-based methods (Valenzuela-Escárcega et al., 2015; Bui et al., 2013) and machine learning-based techniques (Zhang et al., 2022; Li et al., 2010b). While these methods demonstrated good performance, their generalisability and applicability beyond specific domains or annotated datasets have been limited.

Furthermore, EE has been approached as a question-answering (QA) (Du & Cardie, 2020; Pan et al., 2021) or machine reading comprehension (MRC) task (Liu et al., 2020b), fundamentally seeking to answer the "who did what, when, where, why, and how" question as described by Hamborg et al. (2019). The emergence of QA methods has revolutionised web search Kodra & Meçe (2017) and are valuable in clinical medicine, where physicians rely on EHRs to find answers to patient-related inquiries, aiding their clinical decision-making process (Demner-Fushman et al., 2009). This task, known as Extractive QA, has seen significant advancements with models trained on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), a widely used benchmark dataset for extractive question answering, surpassing human performance (Lewis et al., 2019), without necessarily requiring annotations.

In the realm of clinical QA systems, the predominant focus has been on creating specialised search engines tailored to physicians, as evidenced by previous research (Goodwin & Harabagiu, 2016; Cao et al., 2011; Cairns et al., 2011), with commonly used corpora like PubMedQA (Jin et al., 2019). These systems excel in retrieving answers from a knowledge base composed of documents, allowing for the extraction of answers from multiple sources.

Nonetheless, in the context of developing a fall registry, the emphasis shifts towards information contained within individual documents, as this information is inherently patient-specific. Buonocore et al. (2023) successfully developed a QA system for constructing a cardiological registry, surpassing the performance of previous rule-based methods by leveraging their own annotated dataset for fine-tuning.

Considering the novelty of the task of extracting fall mechanisms and the absence of annotated data for fine-tuning models, this thesis only scratches the surface of QA-

based mechanism extraction by leveraging QA models pre-trained on datasets such as SQuAD and PubMedQA. Unlike conventional QA tasks, these models are tailored to address one specific, overarching causal question: "How did the patient fall?"

Chapter 3

Data

This chapter offers an overview of the data pipeline to build the datasets for modelling. Therefore, it excludes any task-specific data transformation processes.

Section 3.1 describes the databases from which the data were collected as well as the method of how the data were collected, including the requirements for our study cohort. Section 3.2 outlines the pre-processing phase, which consists of de-identification, segmentation and cleaning. Section 3.3 provides a detailed description of the down-sampling of the training dataset.

3.1 Data Collection

The data utilised for this study was sourced from the Research Patient Data Registry (RPDR), a centralised clinical database that houses electronic health records from Mass General Brigham, a network of hospitals and healthcare institutions in Massachusetts, United States (Nalichowski et al., 2006; Murphy & Chueh, 2002).

For the training set, hospital and emergency discharge summaries from Massachusetts General Hospital (MGH) and Brigham's Women's Hospital (BWH), two major healthcare facilities connected to the RPDR, were collected. This note type was selected for its descriptive nature and ability to provide contextual insights regarding the patient's clinical journey and reasons for admission. While other note types were considered, they were not included due to the limited availability of retrievable records. This provisional dataset comprised a total of 152,011 discharge summaries written between 2001 and 2019, as well as metadata including corresponding Enterprise Master Patient Index (EMPI), also known as patient ID, and medical record numbers (MRN).

As this thesis focuses on patients who have experienced hip fractures, only patients adhering to specific criteria were selected for inclusion in the evaluation set. These criteria entailed patients being above 18 years of age, undergoing surgical treatment for hip fractures at either MGH or BWH and receiving an ICD-9/ICD-10 diagnosis code corresponding to hip fracture within a maximum window of 30 days, spanning from 2010 to 2018. From the RPDR, a sample of 1000 patients was retrieved. For each patient, the most informative clinical note was manually collected by a medical doctor (MD) and a medical student using EpicCare EMR (Epic Systems Corporation), a software that stores electronic health records. The use of EpicCare was preferred over RPDR due to the limited availability of various note types that are essential for creating a test set that closely represents real-world applications.

As patients from RPDR could also be present in EpicCare, the datasets were compared to identify overlapping entries. Rather than excluding all RPDR patients from the EpicCare dataset, only cases with intersecting clinical notes were removed. This decision was made considering that a patient could have diverse notes in their medical record and resulted in only one overlapping note. Moreover, an additional patient was excluded as the clinical note could not be located in EpicCare. The resulting test set comprised 998 clinical notes, encompassing various note types such as discharge summaries and history and physical examination notes.

3.2 Data Cleaning & Pre-processing

Clinical notes obtained from RPDR and EpicCare EMR are formatted as digital forms or form-based interfaces mapped into plain text. These notes exhibited various characteristics, including designated sections for data entry, the presence of newlines indicative of document structure, and the use of underscores resembling spaces for signing physical documents. Given these unique characteristics, clinical notes require specific preprocessing steps that differ from regular texts.

First, a de-identification process was incorporated into the pipeline to eliminate any personally identifiable information present within the clinical notes, as described

in further detail in section 3.2.1, to safeguard patient privacy and enhance dataset reusability. The de-identified clinical notes were then segmented into distinct sections, utilising our custom parser which has been developed through iterated evaluations of sampled clinical notes. For each of the sections, a sequence of cleaning and filtering techniques was applied for efficient elimination of extraneous data within interconnected clusters of information, leveraging the patterns identified during the evaluation of the segmentation process.

3.2.1 De-identification of Clinical Notes

The de-identification process for the clinical notes in both the training and test sets involved two phases: 1) identification of PHI identifiers, and 2) substitution of these identifiers with entity group names.

To identify any of the 18 PHI identifiers specified in appendix A, we utilised the stanford-deidentifier-base-model, an open-source transformer-based model developed by Chambon et al. (2022) and available on Hugging Face.[1] This model was selected for this thesis because of its performance on de-identifying notes of varying formats, as evidenced by its high F1 score of 98.9 on the de-identification of the I2b2 2014 test set Stubbs & Uzuner (2015).

The Stanford de-identification model was implemented as a Named Entity Recognition (NER) pipeline with a "simple" aggregation strategy to group the entities in the predictions. After an iterative process of evaluating different threshold settings (0.70, 0.80, 0.90, 0.95) and conducting a manual examination of multiple instances, a classification threshold of 0.80 was selected. This threshold effectively identified all entities that required de-identification in the inspected sample set while minimising the occurrence of false positive identifications with incorrect entity names.

Chambon et al. (2022) developed a rule-based post-processor to substitute each PHI span with a synthetic PHI. For example, a commonly used hospital name in the original data would be replaced by a synthetic hospital name. To avoid any biases from the synthetic PHI, we replaced the PHI entities with their respective entity names enclosed

[1] Hugging Face: Stanford AIML

in square brackets, indicating their de-identification. This process involved utilising the start and end indices of each entity and replacing the corresponding text with the detected entity name in reversed order. Additionally, we modified the entity name "HCW" to "HEALTHCARE WORKER" for clarity.

The evaluation of the de-identification process on our dataset was excluded due to its scope extending beyond that of this thesis and time constraints.

3.2.2 Section parsing and development

The section parser was developed in an iterative approach, in which a sample of sectioned clinical notes was analysed and evaluated in each iteration to identify patterns. The parsing algorithm was partially set up as finite-state-machines (FSM), utilising the patterns from each evaluation as transition cases (Cho et al., 2003).

In the initial phase, outlined in Algorithm 1 in Appendix B, each clinical note was segmented based on the newline characters that resulted from the original digital format. For each non-empty segment, the parser checked whether it contained any of the patterns indicating a section heading (Table B), similar to (Cho et al., 2003). Consecutive segments were then concatenated together until the parser encountered a new section heading or reached the end of the input.

A subset of 10 segmented clinical notes of varying types was manually evaluated from the first parse. The evaluation primarily focused on grammatical accuracy, meaning that each section is correctly parsed when there is no inclusion of extraneous words from other sections or omission of any words. This choice was necessitated by a diverse range of note formats, which included the presence or absence of subsections. While the parser demonstrated an accuracy of 0.85 over 781 sections, the analysis identified two main parsing issues: 1) the inclusion of word(s) belonging to the preceding section, which is incorrect in all cases, and 2) the occurrence of multiple sections being encompassed within a single section, deemed accurate when those sections are linguistically correct. Based on these issues, two additional aspects were added to the segmentation process.

Leveraging the sections obtained from the initial heading-body parsing, these seg-

ments were subjected to a secondary FSM. This FSM determined whether the present segment pertained to the preceding section, following the patterns detailed in Table B.2, in a similar fashion as the initial FSM. This setup allowed for the concatenation of a section possessing partial relevance to the previous section, potentially leading to the creation of additional "multiple sections". Therefore, a "multiple sections" parse followed afterwards.

Endeavours to separate the multiple sections contained in one segment posed challenges due to ambiguous section headings (e.g. lacking colons), section bodies without clear conclusions (e.g. absent periods), as well as varying heading levels. Since grammatically accurate sections outweighed the necessity for section separation, the conditions listed in Table B.3 were formulated to enhance precision rather than recall.

Given that these conditions embody concepts rather than specific implementations of regular expressions, we conducted experiments across two iterations using various regular expressions. The first complete parser achieved a score of 0.93, which improved to 0.98 in the second iteration, as shown in Table 3.1. Across both iterations, a noticeable reduction in the instances of "belongs to previous sections" was observed, leading to enhanced accuracy. Despite the presence of segments with multiple sections, their occurrence diminished compared to the exclusive section heading-body parsing approach.

The last iteration of the parser was utilised to parse the clinical notes into sections, of which each section was tagged with its original text identification number to facilitate the concatenation of the sections into their respective full texts.

Parser Components	**Accuracy**
Section heading/body (I)	0.85
Section heading/body + belongs to prior section/multiple sections (II)	0.93
Section heading/body + belongs to prior section/multiple sections (III)	0.98

Table 3.1: Evaluation of a subset of 10 sectioned clinical notes of each iteration based on linguistic accuracy.

3.2.3 Text Cleaning

The following cleaning processes were applied to the sections of the clinical notes:

Filtering identical sections. Discharge summaries often featured sections that were identical as they contained various encounters over the years including similar segments. These duplicated sections do not correlate with significance and were omitted to avoid skewed prediction results in the modelling phase.

Elimination of notifications. Within the dataset, certain texts commenced with notifications indicating that they were converted from PDF files, and potential inaccuracies might exist. Since these notifications were not intrinsic to the summaries themselves, they were removed from the dataset.

Removal of hexadecimal/unicode sections. A portion of the dataset contained sections encoded in unicode or hexadecimal strings, rendering them unreadable. These sections exhibit recognisable characteristics, such as sequences of isolated capital letters interspersed with spaces. However, due to the presence of additional irregular patterns, determining their precise start and end points posed challenges. Therefore, sections featuring the mentioned pattern were fully removed from the dataset using regular expressions to identify the indicated pattern.

Omission of timestamps. Discharge summaries primarily serve as handover documents for healthcare workers and contain detailed logs regarding, for example, specific medical interventions or administered doses. The timestamps associated with these procedures, albeit relevant for healthcare workers, are abundant information that does not provide relevant information for our modelling tasks. Therefore, these timestamps were removed from the sections.

Omission of de-identified information. De-identified information encompassing personal names, IDs, and hospital names transformed into the generic format "[ROLE]". Because of the transformation of personal identifiers to this generic format, the resulting texts are akin to an empty shell. Therefore, words marked as de-identified (e.g. "[PATIENT]") are removed from the texts. This process extended to their combinations with specific prepositions, preventing potential grammatical inconsistencies (e.g. "at the [HOSPITAL]").

Exclusion of administrative sections. Administrative sections encompassed information about document signatories, indicated by the term "signed", as well as instances of the term "FINAL", denoting the finalisation of discharge summary documents. These administrative sections were removed as they were not relevant to the patient.

Discarding sections with uninformative content. Sections containing no substantive content, such as sections containing solely the heading, were removed due to the lack of information.

3.3 Data Downsampling

After performing the aforementioned data cleaning and pre-processing steps, the training dataset of discharge summaries initially comprised 148,421 clinical notes.

However, initial experiments conducted on the complete dataset had to be modified due to computational limitations. Consequently, the experiments were reconfigured to utilise a smaller training dataset.

To reduce the size of the training dataset, a systematic approach was adopted based on the yearly distribution of clinical notes. Specifically, years characterised by the lowest note volumes were excluded. Subsequently, within the retained clinical notes spanning from 2010 to 2018, a straightforward downsampling strategy was applied. This strategy involved the random exclusion of clinical notes that did not contain any synonym of the word "fall", as this represented the majority of the notes. The adjusted training set was ultimately composed of 110,642 discharge summaries.

Chapter 4

Annotation Study & Guideline Development

This chapter describes the annotation study that has been conducted to address the scarcity of fall-related annotation guidelines for NLP tasks. First, the difficulty of each of the three tasks was explored: the identification of fall occurrences, a task that has already been tackled by previous studies (Tremblay et al., 2009; Fu et al., 2022; Patterson et al., 2019), and two novel tasks — fall mechanism extraction and fall impact classification. The difficulty was estimated by comparing annotations from both domain experts and an individual without a medical background in order to yield valuable insights that can substantially improve annotation guidelines, aid in the development of machine learning models, and provide a deeper understanding of the task's feasibility within the broader research context. As a product of the annotation study, the evaluation dataset was developed.

Section 4.1 outlines the methodology and setup of the annotation study. Subsequently, in Section 4.2, the findings from the annotation study are presented, accompanied by a detailed analysis of the specific cases that posed challenges. At last, Section 4.3 concludes the study and outlines the future usage of the annotations throughout the remainder of this thesis.

4.1 Study Set-up

The annotation process for the clinical notes of the study cohort consisted of two stages: 1) initial guideline development and annotation conducted by a medical doctor and a medical student, and 2) annotation conducted by a non-medical Linguistics graduate.

In the first phase, which can be regarded as a pre-study for this study, an extensive review of existing literature on falls and fall-related injuries was conducted by a medical doctor and a medical student. Drawing upon these literature studies, an initial protocol was developed including the inclusion and exclusion criteria to identify fall occurrences, mechanisms, and impact, as detailed in Appendix C and shown in Table 4.1. Time limitations necessitated the distribution of clinical notes between the two annotators, with each annotator responsible for annotating 499 clinical notes. To uphold uniformity and resolve any discrepancies or uncertainties, frequent evaluation meetings were held between the two annotators, leading to refinements in the annotations within the evaluation dataset.

In the second phase, a subset of 15 clinical notes was annotated by a non-medical Linguistics graduate [1] to assess several aspects, including the clarity of the annotation guidelines established by medical annotators. This evaluation aimed to gauge the uniform understanding among annotators, the reproducibility of the annotation task, and the necessary level of clinical expertise.

Cohen's kappa (Cohen, 1960), one of the most commonly used statistics to measure the agreement between annotators while taking into account the possibility of chance agreement (McHugh, 2012), was used to calculate the inter-annotator agreement (IAA) between the annotators with medical backgrounds and the non-medical graduate.

[1] Author of this thesis: J.M. Chan, MA.

Task	Description
Fall Occurrence	Annotators classify instances as falls (1) or non-falls (0) based on the provided definition: "an unintentional change in position resulting in coming to rest on the ground or another lower level". Certain falls, such as those from bikes or due to self-harm, are excluded.
Fall Mechanism	Annotators identify and mark exact words or phrases indicating the method of falling in the note.
Fall Impact	Annotators classify falls as high-impact (1) or low-impact (0) based on specific criteria. High-impact falls (over 1 meter) include explicit mentions of height, falls from high settings, and falls off playgrounds or trampolines. Low-impact falls (less than 1 meter) include slips, falls from low surfaces, and single-step falls. Falls without specific height references are considered low-impact.

Table 4.1: Overview of the initial guidelines per task.

4.2 Results & Discussion

Table 4.2 shows the inter-annotator agreement between the non-medical annotator and the two medical annotators for each of the tasks.

4.2.1 Fall Occurrence

For the detection of fall occurrence, both the medical annotators and the non-medical annotator achieved 0.76 Cohen's kappa, a substantial agreement according to the interpretation scale by Landis & Koch (1977). Out of the 15 annotated notes, only one

disagreement emerged concerning the following section: "a 20 foot ball landing on both lower extremities". This discrepancy can be attributed to the ambiguity of the text and possible typos, as it can be interpreted as either a 20-foot fall (with 'f' instead of the 'b') or 20 footballs landing on an individual (with an additional 'a' and missing '*s*') which does not align with the defined criteria for a fall. The task was generally perceived as straightforward, as the majority of the falls were explicit mentions and represented in layman's terms.

4.2.2 Fall Mechanism

Extracting fall mechanisms involved the task of identifying and marking words that indicated how a fall occurred. However, calculating Cohen's kappa directly for annotations containing variable-length sentences presented a challenge as it is typically used for categorical tasks. To assess inter-annotator agreement in this task, the focus was placed on semantic agreement rather than form, specifically examining whether the annotated mechanisms indicated the same cause of injury. The annotators reached an almost perfect inter-annotator agreement of 0.83 Cohen's kappa, but it is important to note that this adjustment to the agreement metric may have underestimated the task's difficulty. Through inspection of the annotations, the word scoping aspect proved to be particularly challenging as annotators often summarised the fall mechanism when the cause of injury was spread across multiple sentences, resulting in changes to the original sentence structure. On a semantic level, the task was also perceived as straightforward as the majority of fall cases were explicitly mentioned (e.g. "He fell off the stairs") and required no medical inference.

4.2.3 Fall Impact

As for identifying fall impact, it became evident during the literature review phase that this task was of higher difficulty than the other two tasks as the definition of low or high impact varied across the literature. The annotation efforts for classifying fall impact resulted in a moderate inter-annotator agreement of 0.56 Cohen's Kappa. Both medical annotators highlighted the lack of relevant information provided in the clinical

notes to determine the impact of a fall. As the impact of a fall can differ significantly for an elderly person compared to a young adult, external factors such as the patient's age and bone strength are vital for the classification process, highlighting the influence of individual factors beyond the manner of falling itself.

Task	IAA (Cohen's Kappa)
Detecting fall occurrence (yes/no)	0.76
Extracting fall mechanism	0.83
Classifying fall impact (high/low)	0.56

Table 4.2: Inter-annotator agreement between medical annotators and a non-medical annotator.

4.3 Conclusion

This annotation study has illuminated that the tasks of recognising fall occurrences and mechanisms were generally perceived as relatively straightforward due to their explicit mentions in clinical notes. A medical background was not always essential to identify whether a fall had transpired and the manner in which it had occurred. This suggests that these tasks can be effectively carried out by individuals without medical expertise, potentially paving the way for the development of larger fall-annotated datasets.

However, determining the impact of falls, whether high or low, posed a more intricate challenge. Identifying the impact category required the utilisation of implicit information, as it involved deducing the severity based on contextual clues. Additionally, both medical annotators highlighted the need for additional information, some of which might not be available in the clinical notes. Future research endeavours could, therefore, explore additional data sources such as CT scans of the fractured bones, to refine the annotation guidelines for fall impact classification which currently remains a task reserved for medical experts.

In light of these findings, a fine-tuned fall protocol was developed as outlined in Appendix D. The revised protocol kept the initial definitions and mainly addressed

the lack of pragmatism observed in the initial guidelines, particularly for annotators without a medical background. Therefore, the finalised version included annotation categories and illustrative examples.

This refined guideline can serve as a comprehensive framework for identifying both fall occurrences and mechanisms, and provide a base for the subsequent task of classifying fall impact. The annotations generated through this study formed the basis for the evaluation dataset used in identifying fall occurrences and extracting fall mechanisms. However, due to the difficulty observed in the classification of fall impact, this particular task was omitted from the modelling phase.

Chapter 5

Fall Occurrence Detection

This chapter describes the first task in the fall pipeline, which is the task of identifying whether a patient fell or not.

Section 5.1 outlines the method of weak labelling the training dataset for the presence of fall indications. Utilising the weakly labelled dataset, machine learning models are trained as described in Section 5.2. Section 5.3 describes the evaluation methods that are used to assess the models. Section 5.4 shows the performances of the models on the evaluation set.

5.1 Weak Supervision

Following the approach conducted by Chan et al. (2022), the training set was weakly annotated utilising a rule-based method that leverages WordNet.

Each of the existing studies evidently utilised the term "fall" for crafting rule-based models (Patterson et al., 2019; Zhu et al., 2017; Tremblay et al., 2009), which was therefore used as the base word for the WordNet strategy. The initial synsets extracted from "fall" yielded 81 synsets. For each of the synsets, a medical student evaluated the relevance of the synset regarding falls, as defined in our protocol (D), based on the provided definitions and examples that were associated with the synset, and labelled it as indicative of falls or not.

The synsets that were annotated as indicative of falls in the initial phase were utilised to obtain the hypernym paths, which refer to the broader category words in which the word "fall" falls. For each of these paths, the medical student marked the relevant nodes of each hypernym path from which a list of parent synsets was created.

For each parent synset, all hyponym synsets were retrieved, including the parent synsets, and annotated for their relevance to falls. From the included hyponyms, the lemmas were retrieved to construct the WordNet lexicon.

Regular expressions were utilised to weakly label the data, which included linguistic variations to account for past tenses. The overview of words is shown in Table 5.1.

	Lemma
Initial	*fall*
WordNet	*slip, stumble, wipeout, topple, plummet, dive, trip, tumble*

Table 5.1: Breakdown of terms utilised for weak labelling the training data per source.

5.2 Modelling

For the identification of fall occurrences, a variety of classifiers mentioned in previous studies were utilised: rule-based, traditional machine learning classifiers, and BERT.

5.2.1 Rule-based Models

Rule-based methodologies often serve as a foundational benchmark for machine learning models and can be useful for their simplicity, transparency, and ease of manageability (Fu et al., 2022).

In this thesis, five rule-based models were implemented for comparison. The baseline model, classified all notes as positive for fall, whereas the other rule-based models used regular expressions of the selected terms and their linguistic variations. The regular expressions only included complete words to avoid false positives such as "*fell*ow". The second rule-based classifier consisted of linguistic variations of the term "fall", similar to the approach applied by Zhu et al. (2017), and excluded synonyms. The other models included the regular expressions provided by Fu et al. (2022), the WordNet approach, as well as a combination of the two.

5.2.2 Machine Learning

Machine learning models trained in this thesis followed existing work such as Tremblay et al. (2009), McCart et al. (2013), and Tohira et al. (2022). Three distinct classifier models were implemented: logistic regression, support vector machine with a linear kernel, and random forest.

As input for these classifiers, the clinical notes were encoded as document-term matrices using tf-idf.

5.2.3 BERT

BERT is a pre-trained language representation model with exceptional capacity to capture contextual nuances and state-of-the-art performances in recent NLP tasks (Devlin et al., 2019). Following existing work by Fu et al. (2022), a base BERT was implemented with parameters that were adjusted to our computational limitations. A maximum sequence length of 512 was utilised, following the original paper by Devlin et al. (2019), a batch size of 8, and a number of 3 epochs.

5.3 Evaluation

The rule-based classifiers, weakly trained machine learning classifiers, and BERT models were evaluated for their effectiveness in classifying fall occurrence in the hip fracture dataset, 80% of the 998 hip fractures were marked as fall-induced by the medical annotators in Chapter 4. The assessment of their classification capabilities employed established metrics, including precision, recall, and F1 score, aligning with prior research standards (Cusick et al., 2021; Fu et al., 2022; Topaz et al., 2019).

Apart from evaluating different NLP methods, these classifiers were also compared with the existing ICD-based methodology, acting as a reference point to measure the effectiveness of the current operational approach and the potential of NLP to enhance it. This comparative analysis entailed checking the presence of fall-related ICD-codes, detailed in Table E in Appendix E, against the additional falls identified by the NLP methods.

5.4 Results

Table 5.2 shows the performances of a baseline which classifies all notes as a fall. Moreover, it shows the performances of the rule-based classifiers, which consists of the simplest classifier of only the terms "fall" and "fell", the literature- and expert-based rules by Fu et al. (2022), our WordNet approach, and a combined strategy of the aforementioned classifiers. Each rule-based classifier achieved a F1 score of 0.96, outperforming the baseline which achieved a F1 score of 0.89.

Rule-based	**Precision**	**Recall**	**F1**
Baseline	0.81	1.00	0.89
Fall/fell	0.94	0.97	**0.96**
Fu et al. (2022)	0.93	0.99	**0.96**
WordNet	0.94	0.97	**0.96**
Combined	0.93	0.99	**0.96**

Table 5.2: Performances of rule-based classifiers on identifying fall occurrences.

The performances achieved by the machine learning classifiers and BERT, respectively trained and fine-tuned on the weakly labelled training dataset, are shown in Table 5.3. It can be seen that all machine learning models outperformed the baseline, but achieved slightly lower F1 scores than the rule-based classifiers. In this category, the SVM classifier achieved the highest F1 score of 0.95, followed by BERT, Logistic Regression, and Random Forest.

ML/BERT	Precision	Recall	F1	ROC-AUC
Baseline	0.81	1.00	0.89	-
Logistic Regression	0.90	0.97	0.93	0.89
Support Vector Machine	0.93	0.98	**0.95**	**0.92**
Random Forest	0.90	0.94	0.92	0.80
BERT	0.91	0.97	0.94	0.86

Table 5.3: Performances of machine learning classifiers and BERT on the task of identifying fall occurrences.

Among the cohort of 998 patients, approximately 80% (808 patients) suffered a hip fracture related to a fall. However, only 28% (231 patients) of these individuals received an ICD code specifically related to a fall. Table 5.4 shows the percentage of correctly identified fall patients by the NLP models that were missed by ICD codes. It can be seen that the best-performing rule-based classifiers and the SVM were able to detect an additional 70% of fall-related hip fracture patients, resulting in a fall catch rate of 98% in this study cohort. These results show that both rule-based classifiers, as well as more advanced models, can catch fall patients who were missed by the ICD framework.

Method	Falls Caught
Fall/fell	0.69
Fu et al. (2022)	0.70
WordNet	0.70
Combined	0.70
Logistic Regression	0.68
Support Vector Machine	0.70
Random Forest	0.67
BERT	0.69

Table 5.4: Evaluation of NLP-models in their ability to identify patients who experienced falls but did not have corresponding ICD-codes in their medical record.

5.5 Discussion

The task of fall identification has consistently exhibited strong performance in previous research endeavours, a trend reaffirmed by our annotation study. Our findings align with this established trend, underscoring the seemingly uncomplicated nature of fall identification. However, it is crucial to note that our evaluation dataset, predominantly composed of hip fracture patients, exhibited an imbalanced distribution of fall-positive cases (80%), which can be seen in the high-scoring baseline performance. Nonetheless, both rule-based classifiers and advanced machine learning models showcased remarkable results, surpassing the baseline.

Yet, even the leading classifiers, particularly the rule-based ones, faced typical challenges inherent to this category: words not included in predefined rules went unnoticed. This encompassed abbreviations such as "GLF" (ground-level fall), specific instances such as "last off step ladder," and less common fall variations such as "found down".

Conversely, the most effective machine learning classifier, the SVM classifier, successfully identified the note containing "last off step ladder," indicating the SVM had learned patterns related to ladders and falls. Another note is characterised as "No reported fall from nursing home, but the patient reports having fallen in the past. Very unclear" was similarly identified possibly due to the patterns of nursing home and falls. Although the SVM classifier identified a few more specific cases, it also missed certain prototypical fall examples, such as "The patient is a female who tripped over a broom while working" and "fracture after a fall going down stairs".

While weakly supervised ML classifiers performed slightly worse than rule-based models at the current stage, their ability to detect instances beyond standard fall vocabulary demonstrates the potential for their use. This potential could be harnessed by acquiring more annotated data, feasibly done through layman annotations, for training supervised ML models, or by developing a hybrid model consisting of an ML classifier and post-rules for increased generalisability. Another viable strategy is active learning, where current weakly supervised models predict a sample of held-out clinical notes with varying confidence scores, and subsequently correct and incorporate these instances into the training data.

In comparison to the ICD-based approach, the power of NLP becomes strikingly evident. Both the top rule-based and ML classifiers identified approximately 3.5 times more patients who had experienced falls. This stark contrast underscores the unparalleled potential of NLP for clinicians in developing a fall registry, which not only enhances our understanding of fall occurrences but also paves the way for more accurate and expansive healthcare data analysis, thereby transforming the landscape of patient care and safety.

Chapter 6

Fall Mechanism Extraction

This chapter describes the second step of the fall pipeline, which is the task of identifying how someone fell by extracting the words that describe it.

Section 6.1 outlines the pre-trained QA models used to extract fall mechanisms. Section 6.2 describes the evaluation methods to measure the models' extraction capabilities. Section 6.3 shows the performances of each of the models on the evaluation set.

6.1 Modelling

For the novel task of extracting fall mechanisms, this thesis experiments with QA models to identify how someone fell. In contrast to general QA applications, in which a model is expected to answer a variety of questions, only the question "How did the person fall, slip or trip?" was asked to extract fall mechanisms.

This thesis focused specifically on extractive QA models, which, given a context and a question, provide an answer directly extracted from the context. In this case, the context refers to a single clinical note, and the question dictates the specific information to be extracted. The answer generated by the QA model represents the text span from the context identified as the answer to the question. This method minimises the risk of providing irrelevant information.

The choice of pre-trained QA models was made based on their availability and the alignment of their training data sources with the context of our study. The majority of QA models are typically pre-trained on datasets such as SQuAD, which comprises questions related to Wikipedia articles, with answers being specific segments of text

from the corresponding reading passages (Rajpurkar et al., 2016). Given the medical context of our data, a preference was also given for models pre-trained on medical information sources such as PubMedQA, a distinctive biomedical question-answering dataset sourced from PubMed abstracts (Jin et al., 2019).

In light of the aforementioned considerations, four distinct models were chosen for our experiments. The first model in our selection was the standard BERT model (Devlin et al., 2019). BERT is pre-trained on Wikipedia texts for its language modelling component and on SQuAD for its QA functionality. Additionally, RoBERTa, a variant similar to BERT but pre-trained on broader corpora, including CCNews (Liu et al., 2019) in addition to Wikipedia, was included. RoBERTa was also pre-trained on SQuAD for QA tasks.

For models specifically tailored to the medical domain, BioBERT was leveraged. BioBERT is a language model pre-trained on PubMed abstracts and PMC articles, and it is pre-trained on both SQuAD and PubMedQA for QA tasks. Lastly, SapBERT-PubMedBERT was leveraged, a language model pre-trained on UMLS and PubMed abstracts and articles. SapBERT-PubMedBERT is also pre-trained on SQuAD for QA-tasks (Liu et al., 2020a). These models were chosen to provide a comprehensive representation, encompassing both general and medical-specific pre-training data, aligning with the diverse contexts of our study.

In the task of identifying fall occurrences, weak supervision was employed to annotate clinical notes. However, labelling specific word spans to identify mechanisms posed a substantial challenge due to the difficulty of identifying patterns within the texts and the lack of existing studies that delved into this. Considering the novelty of the task of extracting fall mechanisms and the exploratory aim of this study, this thesis only considered pre-trained models without fine-tuning to assess the potential of QA models within this context. Therefore, these pre-trained models were directly applied to the clinical notes in the hip fracture dataset.

6.2 Evaluation

For the extraction of fall mechanisms, only the clinical notes identified with fall occurrences, per the first task, were included. These were 808 of the 998 clinical notes, annotated by the medical experts.

In order to evaluate how well the pre-trained models perform on the extraction of fall mechanisms, three metrics were implemented with varying degrees of strictness: exact matching, F1 score and accuracy.

In the context of exact matching, the response generated by the model must precisely correspond to the annotated answer. In other words, the answer must align word-for-word, adhering to a stringent metric that allows no leeway for variations or flexible spans.

The annotation study indicated that exact matches between annotators were difficult to achieve, and as our interest in fall mechanisms does not lie in the exact span of words, but rather that the mechanism on how someone fell is clear, other evaluation metrics that are more lenient are considered more valuable.

The F1 score, when applied to QA models, quantifies the degree of overlap between the extracted answer and the gold standard. Specifically, this metric relies on the count of common words between the prediction and the gold standard: precision represents the proportion of shared words relative to the total number of words in the prediction, while recall signifies the proportion of shared words in relation to the total number of words in the gold standard.

The accuracy metric, when applied to QA models, considers any word overlap with the gold standard. This means that a perfect score on the accuracy metric can be achieved while F-1 and EM would penalise the instances in which there is no full or considerable overlap. Hence, this metric is more reflective of the experience of the end-user since in many use cases, the context around the predicted answer will also be provided to the user.

Similar to the first task, the NLP methods were compared to the existing ICD-based methodology, acting as a reference point to measure the effectiveness of the current operational approach and the potential of NLP to enhance it. First, the aforementioned

metrics were used to evaluate the corresponding description of the fall-related ICD codes, as detailed in Table E in Appendix E, against the gold standard annotations. Additionally, a comparison is made by checking the presence of fall-related ICD codes against the additional highly probable extracted mechanisms, i.e. with a high degree of overlap with the gold standard, identified by NLP methods.

6.3 Results

Table 6.1 presents the performance results of the ICD code descriptions and various pre-trained models on the extraction of fall mechanisms from clinical notes of patients with a fall-related injury. The models consisted of BERT-base pre-trained on SQuAD, RoBERTa pre-trained on SQuAD, BioBERT pre-trained on SQuAD and PubMedQA, and SapBERT-PubMedBERT pre-trained on SQuAD.

The ICD method and all QA models attained scores near zero on the exact matching metric, the strictest evaluation criterion, with SapBERT achieving the highest score of 0.06. This was expected as this was also considered a challenge in the annotation phase of the two medical experts, and not considered the most valuable metric for our research aim.

As for the other, more lenient metrics, the results show that all QA models outperform the ICD method, which was expected as ICD codes are formulated in a general manner whereas the extractive QA models retrieve the mechanisms from the clinical notes themselves.

BioBERT yielded the highest F1 score of 0.34, while SapBERT achieved the highest accuracy at 0.69. When considering the alternative evaluation metrics, namely F1 score and accuracy, it becomes evident that QA models pre-trained on medical datasets exhibited marginally superior performance compared to models solely trained on SQuAD.

ICD/QA-Models	Exact Matching	F1	Accuracy
ICD	0.00	0.03	0.15
BERT-base (SQuAD)	0.03	0.27	0.59
RoBERTa (SQuAD)	0.03	0.30	0.64
BioBERT (SQuAD & PubMedQA)	0.05	**0.34**	0.66
SapBERT-PubMedBERT (SQuAD)	0.06	0.32	**0.69**

Table 6.1: The performances of ICD and QA-models pre-trained on SQuAD and/or PubMedQA on the fall mechanism extraction task.

Similar to the first task, approximately 28% of the clinical notes had a corresponding ICD code present. Using the best-performing model, BioBERT, a stringent threshold is employed to retrieve the instances with F1 scores exceeding 0.50, as these are more likely to be semantically similar to the gold standard. This threshold-based approach using a pre-trained model showed that an additional 21% fall mechanisms were identified compared to the use of ICD codes alone.

6.4 Discussion

In the realm of extracting fall mechanisms from clinical notes, pre-trained models displayed sub-optimal results across all assessed metrics. A detailed error analysis sheds light on critical issues that challenge the appropriateness of existing evaluation metrics.

Firstly, our analysis underscored that the stringent requirement for an exact word span match, similar to the criterion used in the annotation study, is not the most suitable evaluation standard. The task emphasises context over exact phrase matching, rendering the exact matching metric overly rigid. For example, the phrases “fell while walking their dog” and “fell while walking” represent different spans despite having similar falling mechanisms. Conversely, the accuracy metric was too lenient, leading to high scores for generic answers such as “fall” without providing meaningful context. While the F1 score was more forgiving than exact matching, it still penalised predictions for missing words, even if the prediction made conceptual sense. A potential solution

could involve adopting a scoring mechanism based on semantic similarity, enabling a better capture of task nuances.

A second crucial insight revolved around the formulation of the question presented to the QA models, specifically, "How did the person fall, slip, or trip?". This question elicited responses lacking sufficient context in some cases. For instance, one of the clinical notes described a patient losing balance while getting dressed and falling on their buttock. While the gold standard included the phrase "they were getting dressed this morning and lost their balance", the majority of the models extracted "fell onto their buttock", and only one extracted "on their buttock". This highlights the necessity for questions that are more context-aware.

Another significant factor involved gold standard annotations. Some mechanism annotations included additional words that provided context relevant to other annotators rather than directly reflecting the text snippet. For example, a QA model accurately identified "stumbled on his shoe and fell on his left side", while the gold annotation simply stated "tripped and fell." This misalignment between annotations and model outputs underscores the need for greater consistency, which can be done by having annotators label the same subset and stricter guidelines for extract mechanisms.

Additionally, clinical notes often contained multiple fall mentions, with sections such as hospitalisation summaries or patient illness histories prevailing. Consequently, gold standard annotations featured, for instance, "she tripped when she tried to get her walker", and later in the note, "This patient had a mechanical fall", both captured by BioBERT. Addressing this challenge might involve more detailed segmentation of notes into sections, each specific to minimise information loss compared to document-based methods, holding potential for enhancing model performance.

Lastly, the pre-trained models were not specifically trained for these types of clinical notes, possibly affecting their performance. The annotation study suggested that laymen could annotate this task, providing an opportunity to build a larger annotated dataset for model fine-tuning to improve performance.

Despite significant room for improvement, pre-trained QA models have surpassed the performance of the existing ICD standard. This suggests that although these pre-

trained models are not yet fully equipped for the task, there is potential for refining their capabilities through fine-tuning. While it is not anticipated that current or future models can perfectly extract mechanisms for all patients, these models can still be valuable for clinicians in developing a fall registry. These models could provide clues regarding the potential location of information on fall mechanisms, operating under the assumption that it could be closely linked to the snippets extracted by QA models, but they could also reduce the workload for clinicians by minimising the number of patients requiring detailed chart review through the use of high threshold settings. Finally, the experiments conducted in this thesis have placed the challenge of extracting fall mechanisms on the agenda, paving the way for the development of potential models that fulfil the required criteria.

Chapter 7

Conclusion

In this thesis, the development of fall registries was explored through a series of tasks, including identifying fall occurrences in clinical notes, extracting their mechanisms, and classifying their impact.

The first objective was to lay the groundwork for fall registry development by assessing the necessary medical expertise and developing annotation guidelines for future studies. Our annotation study compared annotations by a non-medical graduate with those of medical annotators. Results indicated that identifying whether a fall occurred and extracting the mechanism were straightforward tasks, not requiring medical expertise due to their explicit descriptions. However, the study was limited to a subset of 15 clinical notes. Moreover, classifying fall impact proved challenging, as clinical notes lacked sufficient information. A further could assess the difficulty of this task using additional data such as CT scans of fractures, and potentially conduct modelling experiments.

The second objective was to develop a comprehensive fall registry for hip fracture patients, achieved through the first two tasks. In fall occurrence identification, both rule-based and weakly supervised advanced machine learning methods produced exceptional results. NLP methods proved highly effective in identifying falls that eluded conventional ICD codes, capturing 98% of falls among hip fracture patients. However, this cohort represented a group highly susceptible to fall-related injuries, making the task relatively easier due to the high positive rate in the test set. Although future work could expand this approach to broader cohorts, current results already demonstrate that NLP models can significantly facilitate clinicians regarding fall registry develop-

ment.

To enhance the fall registry with detailed fall descriptions, the thesis explored the novel task of extracting fall mechanisms for which pre-trained QA models were utilised. Further efforts are required to improve the performances of QA models, which could involve refining guidelines for stricter rules in extracting word spans as well as developing an annotated dataset for fine-tuning the models specifically for this task. Although the models are not yet up to standard, potentially due to the absence of fine-tuning and inconsistencies in annotations, the QA approach outperformed the current ICD method and extracted an additional 21% of mechanisms for hip fracture patients who experienced falls. These results demonstrate the potential of NLP models to extract additional information about falls, which can help clinicians in their chart reviews even if it is a subset of their clinical cohort. However, most importantly, this study places this novel task of fall mechanism extraction on the map, paving the way for future improved models to be developed.

In conclusion, this thesis has not only laid the groundwork for a sophisticated fall registry pipeline but has also contributed valuable insights and methodologies for the broader field of fall-related research. The combined efforts in annotation, NLP, and innovative task formulation have propelled us closer to a more comprehensive understanding of falls, their mechanics, and their impact on patients, paving the way for more effective prevention and treatment strategies in the future.

Appendix A

HIPAA PHI: List of 18 Identifiers

1. Names;
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geo-codes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Phone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social Security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;

14. Web Universal Resource Locators (URLs);

15. Internet Protocol (IP) address numbers;

16. Biometric identifiers, including finger and voice prints;

17. Full face photographic images and any comparable images; and

18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

Appendix B

Section Segmentation

Condition	**Subcondition 1**	**Subcondition 2**	**Example**
Contains a word followed by colon	Segment contains no period in first 10 characters	First letter is capitalised	"Admission date:"
–	Contains a '.' in the first 10 characters.	Part after '.' should contain at least 1 character	"and the admitting diagnosis is shoulder pain. Check-up date: Saturday."
–	(indicates a trail belonging to the previous section)	–	(The text before the first period is added to the previous section, while the content after becomes a new section heading.)
Consists of all uppercase characters	First character is not a digit (exclude time-related information "7 AM", dosages "500 MG PO", and prescription list "1. VITAMIN D3")	–	"HISTORY & PRESENT ILLNESS"

Table B.1: Conditions for marking a segment as a section heading.

Algorithm 1 InitialParse(Text)

```
Initialise an empty list to store the final list of complete sections.
Initialise the section head as None.
Split the text into parts based on newline characters ('\n').
Remove trailing white spaces and empty segments.
for each segment obtained from the split do
    if segment contains indicators for being section head then
        if segment meets full conditions to be section head then
            Make segment section head
            Set the body state to False
        else if section head is not None then
            Append the segment to the section head.
            Set the body state to True.
        else
            Make segment section head
            Set body to False
        end if
    else
        if section head is None then ▷ # for initial segments without section heads
            Make segment section head
            Set body to False
        else
            Append the segment to section head.
            Set body to True
        end if
    end if
    if body is False then
        Append section head to complete list of sections
    end if
end for
    return Final list of complete sections.
```

Condition	**Description**
'AM' or 'PM' at start of section	'AM' and 'PM' follow a time in all cases due to linguistic correctness. Therefore, presence of these words at the start of a section indicate a cut-off.
Specific endings on '.' in prior section	Majority of the periods indicate the ending of a section or sentence, but prior sections ending with words such as 'Mr.' or 'Pt.' indicate that the current section belongs to previous one.
Prior section ends on punctuations indicating opening	Prior sections ending in punctuations such as '(' or ':', expecting words that follow or closing brackets.
Medication lists	Sections containing medication words such as 'mg'
Lab tests	Sections containing measurements such as 'mol/L'
Itemisation/Enumeration	Consecutive sections containing indicators of itemisation (e.g. bullet point) or enumeration (e.g. '3.')
Coordinating conjunctions and prepositions	Prior sections ending on conjunctions or prepositions that indicate the sentence is not finished (e.g. 'and', 'in')
Rest case without section heading indicators	If the section does not contain any section heading indicators, then it can be added to the prior section.

Table B.2: Transition conditions of the finite-state-machine 'belongs to prior section'. Prior section in these conditions refer to the section preceding the section that is currently parsed.

Condition	**Example**
Wh-questions	What was I seen for?
Clinical note type title	Patient Care Referral Form
Period, closing bracket or consecutive underscores followed by a single-worded section heading or section ending on capitalised word	• given. HPI: • __ SIGNED • given. He
Section headings with empty bodies	Name: Age: 23
Lowercased word followed by all uppercased words at the end	fell MEDICATION LIST

Table B.3: Conditions to split multiple sections into separate sections.

Appendix C

Preliminary Guidelines for Falls

Sirls, E., BSc & Taseh, A., MD

In this study, a fall is defined as an "an unintentional change in position resulting in coming to rest on the ground or another lower level" (Morrison et al., 2013). We divide fall types into two groups: low-energy and high-energy. Low-energy or low-impact falls are defined as falls from standing height or a height of less than 1 meter. For example, this includes falls from standing, walking, slipping, out of bed, off stools, falls down a single step, etc. High-energy or high-impact falls are defined as falls from a height greater than 1 meter. This includes falls off of roofs, playgrounds, trampolines, out of trees, falling down multiple steps, etc. (Lim et al., 2021; Zhu et al., 2020).

This fall classification scheme is widely used and reinforced in the literature (Morrison et al., 2013; Bergström et al., 2008; Kennedy et al., 2001). However, to provide a more objective basis for these parameters, falls from standing height (or <1 meter), are associated with lower injury rates, shorter hospital stays and reduced force upon impact. Meanwhile, those with more severe, high-trauma injuries are more likely to have fallen from a standing height or greater (Bergström et al., 2008; Hsieh et al., 2020a; Kennedy et al., 2001). Additionally, these parameters provide a clear, straightforward way of identifying fall types.

Because of confounding injury, we are excluding falls from bikes, self-harm, violence and animals, falls into water, fire and machinery, those struck by vehicles and those with pathologic (metastatic) fracture (Ekbrand et al., 2020; Shepherd et al., 1990; O'Donnell & Connor, 1996; de Vries et al., 2018). We are also excluding patients with old fractures and with incomplete clinical or radiographic medical records (Zhu et al., 2020).

Appendix D

Annotation Guidelines for Falls

Chan, J. M., MA, Sirls, E., BSc & Taseh, A., MD

Foot and Ankle Research and Innovation Laboratory (FARIL)

The annotation guidelines were developed through a comprehensive literature review, iterative revisions based on practical annotation experience, and evaluation discussions conducted by medical student E. Sirls and medical doctor A. Taseh. The background information on fall definitions, mechanisms, and fall types is written by E. Sirls and evaluated by A. Taseh. The finalisation of the annotation guidelines, including the categorisation of annotation examples and the provision of clear examples [1], was conducted by J.M. Chan, the author of this thesis.

In section D.1, a list of abbreviations commonly utilised in clinical texts is provided. The clinical texts are organised in rows within a spreadsheet. The annotation guidelines for detecting fall incidents are outlined in section D.2, while section D.3 presents the guidelines for extracting fall mechanisms. Furthermore, section D.4 details the guidelines for classifying fall impact.

D.1 List of Abbreviations

Ft feet/foot

(Hip) fx (Hip) fracture

L/R (hip) Left/Right (hip)

MVC Motor Vehicle Crash

[1] In order to ensure privacy, the annotation examples have been modified by the author.

MVA Motor Vehicle Accident

Pt Patient

S/P mechanical fall Status Post (mechanical fall)

D.2 Detecting Fall Ocurrences

In this study, a fall is defined as "*an unintentional change in position resulting in coming to rest on the ground or another lower level*" (Morrison et al., 2013). For this annotation task, annotators have to determine the presence or absence of a fall incident using the aforementioned definition. This task is binary, meaning that annotators will **classify each instance as either a fall ('1') or non-fall ('0').**

D.2.1 Prototypical

Prototypical falls refer to a clear fall event that has occurred.

(a) Explicit mentions of the word "falling":

- *"The patient experienced a fall while walking in the hallway."*
- *"Pt fell from the stairs."*

(b) Terms that meet the criteria for fall, such as:

- *"Patient slipped in the bathroom."*
- *"He tripped over a curb."*
- *"She slid off a bar stool."*

D.2.2 Exclusion Criteria

(a) Only explicit mentions should be included, and falls should not be inferred. An example of a **inference** is:

- *"Patient presented with a fractured hip after an accident."*

(b) Because of confounding injury, we are **excluding falls from bikes, self-harm, violence and animals, falls into water, fire and machinery, those struck by vehicles and those with pathologic (metastatic) fracture** (Ekbrand et al., 2020; Shepherd et al., 1990; O'Donnell & Connor, 1996; de Vries et al., 2018).

- *"Patient slid down bicycle at 12mph."*
- *"Patient was involved in a MVC/MVA."*
- *"Patient has a medical history of cancer and is currently undergoing chemotherapy. They were transferred to this hospital due to an acute pathologic hip fracture. While walking, the patient experienced a sudden sensation of their leg giving way."*

(c) We are also **excluding patients with old fractures and with incomplete clinical or radiographic medical records** (Zhu et al., 2020).

D.3 Extracting Fall Mechanism

Fall mechanism describes the specific method through which trauma directly or indirectly affects the human body (Bahr & Krosshaug, 2005; Toney-Butler & Varacallo, 2022). It aims to answer the question: "How did this person fall?"

To annotate fall mechanisms, annotators should aim to identify and **mark the precise words or phrases in the clinical notes that indicate the method or manner of falling**. These identified words should be added to a designated column in the provided spreadsheet.

D.3.1 Prototypical

Prototypical examples, with the annotated fall mechanism indicated within squared brackets:

- *"Fell in the morning after [tripping over a curb]."*
- *"Pt [tripped on a object]."*

- *"She [tripped over the rug], falling on to R side."*

D.3.2 Multiple Sentences

Annotators may encounter challenges when the fall mechanism is described across multiple sentences. In such cases, it is important to capture the relevant information and context accurately. Adjustments can be made to the sentence structure or a summary of the fall mechanism can be provided. **Examples of multiple-sentence mechanisms** are:

- *"Walked down the stairs when he lost his balance and fell down 6 steps."*
- *"A XX-year old male was replacing roof tiles. He fell approximately 2m off of his ladder landing on the right side."*)

Annotators should ensure that they capture the relevant information and context accurately. **In cases of multiple-sentence mechanisms, sentence structure can be adjusted or the fall mechanism can be summarised (aiming to maintain the original text's structure as closely as possible)**:

- *"Walked down the stairs, fell down 6 steps"*
- *"Fell 2m off ladder while replacing roof tiles"*

D.4 Classifying Fall Impact

Falls from standing height (or below 1 meter), are associated with lower injury rates, shorter hospital stays and reduced force upon impact. Meanwhile, those with more severe, high-trauma injuries are more likely to have fallen from a standing height or greater (Bergström et al., 2008; Hsieh et al., 2020b; Kennedy et al., 2001). Based on these parameters, fall types can be classified into two groups: low-energy and high-energy. Low-energy falls are defined as falls from standing height or a height of less than 1 meter. High-energy falls are defined as falls from a height greater than 1 meter. This classification scheme for falls is widely used and reinforced in the literature (Morrison et al., 2013; Bergström et al., 2008; Hsieh et al., 2020b).

For this annotation task, annotators have to classify whether a fall is low-impact or high-impact. This task is binary, meaning that annotators will **classify each text as either a high ('1') or low ('0') impact.**

D.4.1 High Energy

(a) **Explicit mentions of the height (above 1 meter)** serve as prototypical examples that fall within the definition of high-energy falls:

- *"Patient suffered injuries after falling from a 20ft tree."*

(b) Falls occurring in **high height settings, such as roofs and stairs**, and **standing on elevated surfaces**, can typically be considered higher than 1 meter, even without explicit mention:

- *"Pt slipped and fell of the roof."*
- *"She fell down the stairs."*
- *"Patient was changing a light bulb, standing on a bar stool, when she slipped."*

(c) Clinical texts may contain **explicit mentions of specific measurements or general indications of a fall from a significant height**. Examples include:

- "*He tumbled down two flights of stairs.*"
- *"They fell down multiple steps."*

(d) According to the literature (Lim et al., 2021; Zhu et al., 2020), **falls off playgrounds and trampolines** can also be considered high-energy falls.

D.4.2 Low Energy

(a) **Falls from standing height, including slipping**, are considered prototypical low-energy falls:

- *"She tripped while walking her dog."*
- *"Slipped getting out of the car."*
- *"Slipped on a slippery floor after getting out of the shower."*

(b) Falls from **low surfaces** can typically be considered lower than 1 meter, even without explicit mention:

- *"She fell out of the bed."*
- *"Patient slid from a bar stool."*

(c) **Single-step falls** can typically be considered lower than 1 meter, even without explicit mention:

- *"Fell after tripping on the last step at home."*
- *"She fell from the stairs and missed the last step and fell."*

(d) If it is **none of the above mentioned categories**, and if there are **no explicit mentions of specific heights** in the note, it can be considered a low-energy fall as the height was not of importance.

- *"S/P Mechanical fall."*
- *"Patient was hospitalised with: FALL."*

Appendix E

ICD codes for Falls

ICD Code	**Description**
W00-W19	Slipping, tripping, stumbling, and falls
Z91.81	At risk for falling
Z91.82	History of falling
R29.6	Repeated falls
E880	Accidental fall on or from stairs or steps
E881	Accidental fall on or from ladders or scaffolding
E882	Accidental fall from or out of a building or other structure
E883	Accidental fall into a hole or other opening in the surface
E884	Other accidental falls from one level to another
E885	Accidental fall on the same level from slipping, tripping, or stumbling
E886	Accidental fall on the same level from collision, pushing, or shoving by or with another person
E887	Other and unspecified falls

Table E.1: ICD Codes for Slipping, Tripping, Stumbling, and Falls

Bibliography

Bahr, R., & Krosshaug, T. (2005). Understanding injury mechanisms: a key component of preventing injuries in sport. *British journal of sports medicine*, *39*(6), 324–329.

Beckwith, B. A., Mahaadevan, R., Balis, U. J., & Kuo, F. (2006). Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC medical informatics and decision making*, *6*, 1–9.

Bergen, G., Stevens, M. R., & Burns, E. R. (2016). Falls and fall injuries among adults aged ≥65 years — united states, 2014. *Morbidity and Mortality Weekly Report*, *65*(37), 993–998.

Bergström, U., Björnstig, U., Stenlund, H., Jonsson, H., & Svensson, O. (2008). Fracture mechanisms and fracture pattern in men and women aged 50 years and older: a study of a 12-year population-based injury register, umeå, sweden. *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, *19*, 1267–73.

Bilimoria, K. Y., Stewart, A. K., Winchester, D. P., & Ko, C. Y. (2008). The national cancer data base: a powerful initiative to improve cancer care in the united states. *Annals of surgical oncology*, *15*, 683–690.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, *32*, 267–270.

Bui, Q.-C., Campos, D., van Mulligen, E., & Kors, J. (2013). A fast rule-based approach for biomedical event extraction. In *proceedings of the BioNLP shared task 2013 workshop*, (pp. 104–108).

Buonocore, T. M., Parimbelli, E., Tibollo, V., Napolitano, C., Priori, S., & Bellazzi, R. (2023). A rule-free approach for cardiological registry filling from italian clinical notes with question answering transformers. In *International Conference on Artificial Intelligence in Medicine*, (pp. 153–162). Springer.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., & Savova, G. K. (2011). The mipacq clinical question answering system. In *AMIA annual symposium proceedings*, vol. 2011, (p. 171). American Medical Informatics Association.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., & Yu, H. (2011). Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, *44*(2), 277–288.

Chambon, P. J., Wu, C., Steinkamp, J. M., Adleberg, J., Cook, T. S., & Langlotz, C. P. (2022). Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *Journal of the American Medical Informatics Association*.

Chan, J.-Z. M., Kunneman, F., Morante, R., Lösch, L., & Zuiderent-Jerak, T. (2022). Leveraging social media as a source for clinical guidelines: A demarcation of experiential knowledge. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, (pp. 203–208). Gyeongju, Republic of Korea: Association for Computational Linguistics.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

Cho, P. S., Taira, R. K., & Kangarloo, H. (2003). Automatic section segmentation of medical reports. In *AMIA Annual Symposium Proceedings*, vol. 2003, (p. 155). American Medical Informatics Association.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cusick, M., Adekkanattu, P., Campion Jr, T. R., Sholle, E. T., Myers, A., Banerjee, S., Alexopoulos, G., Wang, Y., & Pathak, J. (2021). Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of psychiatric research*, *136*, 95–102.

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of statistical planning and inference*, *6*(1), 73–85.

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

de Vries, R., Reininga, I., Pieske, O., Lefering, R., El Moumni, M., & Wendt, K. (2018). Injury mechanisms, patterns and outcomes of older polytrauma patients—an analysis of the dutch trauma registry. *PLOS ONE*, *13*, e0190587.

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L., et al. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, *20*(1), 84–94.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, *42*(5), 760–772.

Denny, J. C., Spickard III, A., Johnson, K. B., Peterson, N. B., Peterson, J. F., & Miller, R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, *16*(6), 806–815.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).

Dos Santos, H. D., Silva, A. P., Maciel, M. C. O., Burin, H. M. V., Urbanetto, J. S., & Vieira, R. (2019). Fall detection in ehr using word embeddings and deep learning. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, (pp. 265–268). IEEE.

Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., & Mark, R. G. (2004). Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*, (pp. 341–344). IEEE.

Du, X., & Cardie, C. (2020). Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 671–683). Online: Association for Computational Linguistics.

Edinger, T., Demner-Fushman, D., Cohen, A. M., Bedrick, S., & Hersh, W. (2017). Evaluation of clinical text segmentation to facilitate cohort retrieval. In *AMIA Annual Symposium Proceedings*, vol. 2017, (p. 660). American Medical Informatics Association.

Ekbrand, H., Ekman, R., Thodelius, C., & Möller, M. (2020). Fall-related injuries for three ages groups – analysis of swedish registry data 1999–2013. *Journal of Safety Research*, *73*, 143–152.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data.

In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 213–220).

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., & Drake, C. (2018). Medical costs of fatal and nonfatal falls in older adults. *Journal of the American Geriatrics Society*, *66*(4), 693–698.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, *1*(2), 161–174.

Fu, S., Thorsteinsdottir, B., Zhang, X., Lopes, G. S., Pagali, S. R., LeBrasseur, N. K., Wen, A., Liu, H., Rocca, W. A., Olson, J. E., et al. (2022). A hybrid model to identify fall occurrence from electronic health records. *International journal of medical informatics*, *162*, 104736.

Gjerstorff, M. L. (2011). The danish cancer registry. *Scandinavian journal of public health*, *39*(7), 42–45.

Gliklich, R. E., Leavy, M. B., & Dreyer, N. A. (2020). Patient registries. In *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 4th edition*. Agency for Healthcare Research and Quality (US).

Gobbel, G. T., Matheny, M. E., Reeves, R. R., Akeroyd, J. M., Turchin, A., Ballantyne, C. M., Petersen, L. A., & Virani, S. S. (2022). Leveraging structured and unstructured electronic health record data to detect reasons for suboptimal statin therapy use in patients with atherosclerotic cardiovascular disease. *American Journal of Preventive Cardiology*, *9*, 100300.

Goodwin, T. R., & Harabagiu, S. M. (2016). Medical question answering for clinical decision support. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, (pp. 297–306).

Griffon, N., Charlet, J., Darmoni, S., et al. (2014). Managing free text for secondary use of health data. *Yearbook of medical informatics*, *23*(01), 167–169.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, *3*(1).
URL https://doi.org/10.1145/3458754

Hamborg, F., Breitinger, C., & Gipp, B. (2019). Giveme5w1h: A universal system for extracting main events from news articles. In *7th International Workshop on News Recommendation and Analytics*, (pp. 35–43).

Hartman, T., Howell, M. D., Dean, J., Hoory, S., Slyper, R., Laish, I., Gilon, O., Vainstein, D., Corrado, G., Chou, K., et al. (2020). Customization scenarios for de-identification of clinical notes. *BMC medical informatics and decision making*, *20*(1), 1–9.

Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U., & Klakow, D. (2020). Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 2580–2591). Online: Association for Computational Linguistics.

Hill, A.-M., Hoffmann, T., Hill, K., Oliver, D., Beer, C., McPhail, S., Brauer, S., & Haines, T. P. (2010). Measuring falls events in acute hospitals—a comparison of three reporting methods to identify missing data in the hospital reporting system. *Journal of the American Geriatrics Society*, *58*(7), 1347–1352.

Hsieh, T.-M., Tsai, C.-H., Liu, H.-T., Huang, C.-Y., Chou, S.-E., Su, W.-T., Hsu, S.-Y., & Hsieh, C.-H. (2020a). Effect of height of fall on mortality in patients with fall accidents: a retrospective cross-sectional study. *International journal of environmental research and public health*, *17*(11), 4163.

Hsieh, T.-M., Tsai, C.-H., Liu, H.-T., Huang, C.-Y., Chou, S.-E., Su, W.-T., Hsu, S.-Y., & Hsieh, C.-H. (2020b). Effect of height of fall on mortality in patients with fall acci-

dents: A retrospective cross-sectional study. *International Journal of Environmental Research and Public Health*, *17*, 4163.

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.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). Mimic-iv-note: Deidentified free-text clinical notes.

Kennedy, R., Grant, P., & Blackwell, D. (2001). Low-impact falls: Demands on a system of trauma management, prediction of outcome, and influence of comorbidities. *The Journal of trauma*, *51*, 717–24.

Khorgami, Z., Fleischer, W. J., Chen, Y.-J. A., Mushtaq, N., Charles, M. S., & Howard, C. A. (2018). Ten-year trends in traumatic injury mechanisms and outcomes: a trauma registry analysis. *The American Journal of Surgery*, *215*(4), 727–734.

Kodra, L., & Meçe, E. K. (2017). Question answering systems: A review on present developments, challenges and trends. *International Journal of Advanced Computer Science and Applications*, *8*(9), 217–224.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, (pp. 159–174).

Leucht, P., Fischer, K., Muhr, G., & Mueller, E. J. (2009). Epidemiology of traumatic spine fractures. *Injury*, *40*(2), 166–172.

Lewis, P., Denoyer, L., & Riedel, S. (2019). Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (pp. 4896–4910).

Li, X.-B., & Qin, J. (2017). Anonymizing and sharing medical text records. *Information Systems Research*, *28*(2), 332–352.

Li, Y., Lipsky Gorman, S., & Elhadad, N. (2010a). Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM international health informatics symposium*, (pp. 744–750).

Li, Z., Liu, F., Antieau, L., Cao, Y., & Yu, H. (2010b). Lancet: a high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, *17*(5), 563–567.

Lim, M. A., Mulyadi Ridia, K. G., & Pranata, R. (2021). Epidemiological pattern of orthopaedic fracture during the covid-19 pandemic: A systematic review and meta-analysis. *Journal of Clinical Orthopaedics and Trauma*, *16*, 16–23.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2020a). Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., Ravikumar, K., Wu, S. T., Kullo, I. J., & Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, *2013*, 149.

Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020b). Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1641–1651).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, *75*, S34–S42.

Ma, M. D., Taylor, A., Wang, W., & Peng, N. (2023). DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), (pp. 15898–15917). Toronto, Canada: Association for Computational Linguistics.

McCart, J. A., Berndt, D. J., Jarman, J., Finch, D. K., & Luther, S. L. (2013). Finding falls in ambulatory care clinical documents using statistical text mining. *Journal of the American Medical Informatics Association*, *20*(5), 906–914.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

McKenzie, K., Harding, L. F., Walker, S. M., Harrison, J. E., Enraght-Moony, E. L., & Waller, G. S. (2006). The quality of cause-of-injury data: where hospital records fall down. *Australian and New Zealand journal of public health*, *30*(6), 509–513.

Meystre, S. M., Ferrández, O., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2014). Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of biomedical informatics*, *50*, 142–150.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Moreland, B., Kakara, R., & Henry, A. (2020). Trends in nonfatal falls and fall-related injuries among adults aged ≥65 years—united states, 2012-2018. *Morbidity and Mortality Weekly Report*, *69*(27), 875.

Morrison, A., Fan, T., Sen, S. S., & Weisenfluh, L. (2013). Epidemiology of falls and osteoporotic fractures: a systematic review. *ClinicoEconomics and outcomes research: CEOR*, *5*, 9.

Mowery, D., Wiebe, J., Visweswaran, S., Harkema, H., & Chapman, W. W. (2012). Building an automated soap classifier for emergency department reports. *Journal of biomedical informatics*, *45*(1), 71–81.

Murphy, S. N., & Chueh, H. C. (2002). A security architecture for query tools used

to access large biomedical databases. In *Proceedings of the AMIA Symposium*, (p. 552). American Medical Informatics Association.

Nalichowski, R., Keogh, D., Chueh, H. C., & Murphy, S. N. (2006). Calculating the benefits of a research patient data repository. In *AMIA annual symposium proceedings*, vol. 2006, (p. 1044). American Medical Informatics Association.

Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., & Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, *8*(1), 1–17.

Norgeot, B., Muenzen, K., Peterson, T. A., Fan, X., Glicksberg, B. S., Schenk, G., Rutenberg, E., Oskotsky, B., Sirota, M., Yazdany, J., et al. (2020). Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, *3*(1), 57.

O'Donnell, C., & Connor, D. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, *28*(6), 739–753.

Olsen, S., Neale, G., Schwab, K., Psaila, B., Patel, T., Chapman, E. J., & Vincent, C. (2007). Hospital staff should use more than one method to detect adverse events and potential adverse events: incident reporting, pharmacist surveillance and local real-time record review may all have a place. *BMJ Quality & Safety*, *16*(1), 40–44.

Palmer, E. L., Hassanpour, S., Higgins, J., Doherty, J. A., & Onega, T. (2019). Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC medical informatics and decision making*, *19*(1), 1–10.

Pan, Q., Chen, X., & Chen, D. (2021). Medical event extraction with question answerability judgment. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, (pp. 1–5). VDE.

Parkin, D. M. (2006). The evolution of the population-based cancer registry. *Nature Reviews Cancer*, *6*(8), 603–612.

Parkkari, J., Kannus, P., Palvanen, M., Natri, A., Vainio, J., Aho, H., Vuori, I., & Järvinen, M. (1999). Majority of hip fractures occur as a result of a fall and impact on the greater trochanter of the femur: a prospective controlled hip fracture study with 206 consecutive patients. *Calcified tissue international*, *65*, 183–187.

Patterson, B., Jacobsohn, G., Shah, M., Song, Y., Maru, A., Venkatesh, A., Zhong, M., Taylor, K., Hamedani, A., & Mendonça, E. (2019). Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department. *BMC Medical Informatics and Decision Making*, *19*.

Pomares-Quimbaya, A., Kreuzthaler, M., & Schulz, S. (2019). Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, *19*, 1–20.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (pp. 2383–2392). Austin, Texas: Association for Computational Linguistics.
URL https://aclanthology.org/D16-1264

Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., & Cavedon, L. (2022). The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, *55*(2), 1–40.

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., & Chute, C. G. (2008). Mayo clinic nlp system for patient smoking status identification. *Journal of the American Medical Informatics Association*, *15*(1), 25–28.

Schmidt, M., Schmidt, S. A. J., Sandegaard, J. L., Ehrenstein, V., Pedersen, L., & Sørensen, H. T. (2015). The danish national patient registry: a review of content, data quality, and research potential. *Clinical epidemiology*, (pp. 449–490).

Shepherd, J. P., Shapland, M., Pearce, N. X., & Scully, C. (1990). Pattern, severity and aetiology of injuries in victims of assault. *Journal of the Royal Society of Medicine*, *83*(2), 75–78.

Shiner, B., Neily, J., Mills, P. D., & Watts, B. V. (2020). Identification of inpatient falls using automated review of text-based medical records. *Journal of Patient Safety*, *16*(3), e174–e178.

Skentzos, S., Shubina, M., Plutzky, J., & Turchin, A. (2011). Structured vs. unstructured: factors affecting adverse drug reaction documentation in an emr repository. In *AMIA annual symposium proceedings*, vol. 2011, (p. 1270). American Medical Informatics Association.

Stubbs, A., & Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, *58*, S20–S29.

Sumrein, B., Huttunen, T., Launonen, A., Berg, H., Felländer-Tsai, L., & Mattila, V. (2017). Proximal humeral fractures in sweden—a registry-based study. *Osteoporosis International*, *28*, 901–907.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, (p. 333). American Medical Informatics Association.

Tan, J. C., Ferdi, A. C., Gillies, M. C., & Watson, S. L. (2019). Clinical registries in ophthalmology. *Ophthalmology*, *126*(5), 655–662.

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). Statistical section segmentation in free-text clinical records. In *Language Resources and Evaluation Conference (LREC)*, (pp. 2001–2008).

Tohira, H., Finn, J., Ball, S., Brink, D., & Buzzacott, P. (2022). Machine learning and natural language processing to identify falls in electronic patient care records from ambulance attendances. *Informatics for Health and Social Care*, *47*(4), 403–413.

Toney-Butler, T. J., & Varacallo, M. (2022). Motor vehicle collisions. In *StatPearls [Internet]*. StatPearls Publishing.

Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, *90*, 103103.

Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R., & French, D. D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, *10*(4), 253–265.

Unguryanu, T. N., Grjibovski, A. M., Trovik, T. A., Ytterstad, B., & Kudryavtsev, A. V. (2020). Mechanisms of accidental fall injuries and involved injury factors: a registry-based study. *Injury epidemiology*, *7*, 1–10.

Uzuner, Ö., 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.

Uzuner, O., Szolovits, P., & Kohane, I. (2006). i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association*. Citeseer.

Valenzuela-Escárcega, M. A., Hahn-Powell, G., Surdeanu, M., & Hicks, T. (2015). A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, (pp. 127–132).

Vollmer, T. L., Ni, W., Stanton, S., & Hadjimichael, O. (1999). The narcoms patient registry: a resource for investigators. *International Journal of MS Care*, *1*(1), 28–34.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. (2018). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, *77*, 34–49.

Workman, T. (2013). Defining patient registries and research networks. *Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks [Internet]. Rockville: Agency for Healthcare Research and Quality*.

Zhang, S., Li, Y., Li, S., & Yan, F. (2022). Bi-lstm-crf network for clinical event extraction with medical knowledge features. *IEEE Access*, *10*, 110100–110109.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, *5*(1), 44–53.

Zhu, V. J., Walker, T. D., Warren, R. W., Jenny, P. B., Meystre, S., & Lenert, L. A. (2017). Identifying falls risk screenings not documented with administrative codes using natural language processing. In *AMIA annual symposium proceedings*, vol. 2017, (p. 1923). American Medical Informatics Association.

Zhu, Y.-B., Chen, W., Xin, X., Yin, Y., Hu, J., Lv, H., xu Li, W., Deng, X., Zhu, C., Zhu, J., Zhang, J., Ye, F., Chen, A.-M., Wu, Z., Ma, Z., Zhang, X., Gao, F., Li, J., Wang, C., Zhang, Y., & Hou, Z. (2020). Epidemiologic characteristics of traumatic fractures in elderly patients during the outbreak of coronavirus disease 2019 in china. *International Orthopaedics*, *44*, 1565–1570.