Unlocking the Potential of Electronic Health Records With Danish Clinical Language Models for Text Mining

Pedersen, Jannik Skyttegaard

DOI:
10.21996/e9cj-9c53

Publication date:
2023

Document version:
Final published version

Citation for pulished version (APA):
Pedersen, J. S. (2023). *Unlocking the Potential of Electronic Health Records With Danish Clinical Language Models for Text Mining*. [Ph.D. thesis, SDU]. Syddansk Universitet. Det Tekniske Fakultet.
https://doi.org/10.21996/e9cj-9c53

Unlocking the Potential of Electronic Health Records With Danish Clinical Language Models for Text Mining

A dissertation submitted for the degree of
Doctor of Philosophy

Jannik Skyttegaard Pedersen

SDU Robotics
The Maersk Mc-Kinney Moller Institute
The Faculty of Engineering
University of Southern Denmark

Odense, August 2023

Supervisor: Thiusius Rajeeth Savarimuthu

SDU Robotics
The Maersk Mc-Kinney Moller Institute
Faculty of Engineering
University of Southern Denmark

Campusvej 55
5230 Odense M, Denmark
Phone +45 65507380
www.sdu.dk

Co-supervisor: Pernille Just Vinholt

Department of Clinical Biochemistry
Odense University Hospital

J. B. Winsløws Vej 4
5000 Odense C, Denmark
Phone +45 66113333
www.ouh.dk

Abstract

Unlocking the Potential of Electronic Health Records With Danish Clinical Language Models for Text Mining

This PhD dissertation focuses on the development of language technology that can be used to extract clinical information from Danish electronic health records (EHRs). EHRs contain important health-related information that can be used to guide the treatment of patients. However, a large part of the information is stored in unstructured narrative text of the EHR, making it difficult and time-consuming to extract the relevant details, especially in acute situations. Consequently, important information may be lost which can increase the risk of misdiagnosis and adverse treatment outcomes.

The recent paradigm shift in the field of natural language processing (NLP), driven by self-supervised neural networks and the transformer architecture, has produced automatic text-processing tools with unprecedented performances. These tools could be used to extract and structure the information from the narrative text of EHRs automatically. However, research in language technology has mostly been explored for high-resource languages like English, while the development of Danish language technology has received less attention, especially for specialized domains such as the clinical.

This dissertation explores the potential of language technology to automatically extract information from the narrative text of Danish EHRs. Moreover, it emphasizes the importance of developing language resources tailored for the Danish clinical domain, as it can be used to enhance clinical research possibilities and improve patient treatment.

The dissertation covers the development of two Danish pre-trained language models which show improved performance compared to existing Danish language models. Moreover, it explores the impact of dataset curation on potential biases in clinical language models. The dissertation also investigates how language models can be used to extract bleeding events from Danish EHRs and evaluates the performance of medical doctors in identifying relevant information when using the bleeding algorithm as an assistive tool. Finally, the dissertation presents a pre-trained language model that can be used to extract clinical information such as diseases, symptoms, and treatments in the narrative text of Danish EHRs.

Resumé

Frigørelse af potentialet for Elektroniske Patientjournaler med Danske Kliniske Sprogmodeller til Tekstmining

Denne afhandling fokuserer på udvikling af sprogteknologi til udtræk af klinisk information fra danske elektroniske patientjournaler. Elektroniske patientjournaler indeholder vigtig sundhedsrelateret information, som kan bruges til at guide behandlingen af patienter. En stor del af informationen i patientjournalen er dog beskrevet i ustruktureret tekst, hvilket gør det vanskeligt og tidskrævende at udtrække relevante detaljer, især i akutte situationer. Som følge heraf kan vigtig information gå tabt, hvilket kan øge risikoen for fejldiagnosticering og forringede behandlingsresultater.

Det nylige paradigmeskifte inden for natural language processing, drevet af self-supervised neurale netværk og transformerarkitekturen, har produceret automatiske tekstbehandlingsværktøjer med hidtil uset præcision. Disse værktøjer kan bruges til at udtrække og strukturere informationen fra den ustrukturerede tekst i den elektroniske patientjournal automatisk. Forskning indenfor sprogteknologi er dog mest blevet udforsket for ressourcestærke sprog såsom engelsk, mens udviklingen indenfor dansk sprogteknologi har været mere stillestående, især for specialiserede domæner såsom det kliniske.

Denne afhandling undersøger potentialet for sprogteknologi til automatisk at udtrække information fra den ustrukturerede del af den elektroniske patientjournal. Derudover beskriver afhandlingen vigtigheden af at udvikle sproglige ressourcer specifikt til det danske kliniske domæne, da det kan bruges til at forbedre behandlingen af patienter samt give nye kliniske forskningsuligheder.

Afhandlingen beskriver udviklingen af to danske præ-trænede sprogmodeller, som viser forbedret præcision sammenlignet med eksisterende danske sprogmodeller. Desuden udforskes det hvordan data curation kan påvirke bias i kliniske sprogmodeller. Afhandlingen undersøger også, hvordan sprogmodeller kan bruges til at udtrække information omkring blødning fra danske elektroniske patientjournaler, og evaluerer lægers evne til at udtrække relevante informationer med blødningsalgoritmen som hjælpeværktøj. Dernæst præsenterer afhandlingen en præ-trænet sprogmodel, som kan bruges til at udtrække kliniske informationer såsom sygdomme, symptomer og behandlinger i den ustrukturerede tekst i danske elektroniske patientjournaler.

Acknowledgements

I would like to thank Thiusius Rajeeth Savarimuthu and Pernille Just Vinholt for your support and friendship during my PhD-studies. Thank you for always being open to exploring new research areas and for sparking my interest in clinical natural language processing. Working with you has been awesome, thank you!

Thank you, Cristina Soguero-Ruiz; for your hospitality, friendship, and academic guidance throughout my research stay in Fuenlabrada. Your insights and positive energy made my visit an absolute pleasure. Gracias!

To my amigo, Martin Sundahl Larsen, working with you has been like hanging out at Manso's on a sunny day. Gracias!

Thank you to all my other colleagues at SDU Robotics and The Department of Clinical Biochemistry at Odense University Hospital. A special thanks to Rasmus Søgaard Hansen, Rasmus Bank Lynggaard, Anne Bryde Alnor, and Signe Westerberg Biørn for your collaboration.

Table of Contents

List of Abbreviations

ADE	**A**dverse **D**rug **E**vents
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
Clin-ELECTRA	**C**linical ELECTRA
CNN	**C**onvolutional **N**eural **N**etwork
DeBERTa	**D**ecoding-**e**nhanced **BERT** with Disentangled **A**ttention
EHR	**E**lectronic **H**ealth **R**ecord
ELECTRA	**E**fficiently **L**earning an **E**ncoder that **C**lassifies **T**oken **R**eplacements **A**ccurately
FNN	**F**eedforward **N**eural **N**etwork
GLUE	**G**eneral **L**anguage **U**nderstanding **E**valuation
GRU	**G**ated **R**ecurrent **U**nit
LSTM	**L**ong **S**hort-**T**erm **M**emory network
MD	**M**edical **D**octor
MeDa-BERT	**M**edical **D**anish **BERT**
MLM	**M**asked **L**anguage **M**odeling
NLP	**N**atural **L**anguage **P**rocessing
NSP	**N**ext **S**entence **P**rediction
OUH	**O**dense **U**niversity **H**ospital
RNN	**R**ecurrent **N**eural **N**etwork
RoBERTa	A **R**obustly **O**ptimized **BERT** pre-training **A**pproach
SOTA	**S**tate-**o**f-**t**he-**a**rt

CHAPTER 1

Introduction

The electronic health record (EHR) is a tool used by healthcare professionals to store information about the patient's disease and treatment history. It includes information such as their medical history, diagnoses, medications, allergies, and lab results. The purpose of the EHR is to provide a detailed overview of the patient's medical information and to share this information among different healthcare professionals involved in the patient's treatment [1]. Healthcare professionals use the information to make informed clinical decisions and secure optimal patient care. The information is stored in a semi-structured way with a large part stored in unstructured narrative text. The narrative text of the EHR allows medical doctors (MD) to communicate complex health information that cannot easily be set into a structured format as each patient can have a unique disease history with different important clinical parameters. However, the amount of information stored in the narrative text makes it complex and time-expensive to obtain the relevant details about a patient, especially in acute situations [2, 3]. As a result, important information can easily be lost causing an increased risk of misdiagnosis and low quality of patient treatment.

Research in natural language processing (NLP), the study of making computers understand human language, could potentially be used to automatically extract and structure information from the narrative text of EHRs. This structured information can be used by MDs to make informed decisions and ultimately improve patient safety. Recently, the field of NLP has gone through a paradigm shift, transitioning from using traditional deep learning-based architectures such as the recurrent neural network (RNN) towards using pre-trained language models. Using self-supervised learning, pre-trained language models can learn to represent and infer the meaning of language simply by being exposed to a large amount of text [4, 5]. Pre-trained language models have achieved state-of-the-art (SOTA) performance on numerous NLP benchmarks, generating considerable interest in language technology across various sectors, including the healthcare sector [6].

In clinical research, language technology could significantly reduce the need for manual data annotation and facilitate insights to be extracted from an unprecedently volume of data [6]. In clinical practice, language technology could be implemented to provide fast and accurate identification of critical information and reduce the time-expensive burden of reviewing the EHR which could increase MDs' availability for direct patient interaction [7, 8]. Language technology can thereby help address some of the challenges currently facing the healthcare sector where an aging population with increased healthcare demands and expectations is pressuring the system [9]. Exploring innovative solutions that can enhance clinical research and improve patient treatment while at the same time reducing institutional healthcare costs is important, and language technology has the potential to play a significant role in addressing this [7].

However, despite of the potential benefits of using language technology in clinical research and practice, it is important to address the potential risks associated with its implementation. As the language models encode knowledge from the text corpus it is exposed to during development, it also naturally encodes any biases present in that corpus [10]. Implementing biased models into clinical practice could cause significant harm, as the model's predictions could have a direct effect on all patients in the institution it is utilized. Dataset curation, the process of creating the datasets used to develop language models, is therefore an important step for avoiding bias issues. This process includes data selection; preprocessing; annotation; and splitting data into train, validation, and test sets. While bias can arise in any of these steps, this dissertation focuses on the bias that can arise due to how samples from patient subgroups are distributed within the datasets used to develop clinical language models.

Development of language technology has primarily been explored for high-resource languages such as English while language technology for the Danish language has gained less attention historically. This is caused by Denmark's relatively small size as a language community and commercial market, and a high proficiency of English [11]. Despite this, there has recently been an increase in the development of open-source Danish language technology [12]. However, the models are primarily developed on, and thereby tailored for, text from the internet, e.g., Wikipedia and online debate forums. This is not ideal for the clinical domain because the language used in EHRs differs significantly from that of the general domain. For example, EHR text contains domain-specific abbreviations, rare medical terminology, and has a high rate of spelling and typing errors [13, 14]. Moreover, the models carry a risk of having embedded biases from the text they have been developed on, e.g., stereotypes and biases from online debate forums.

Development of Danish clinical language technology is, therefore, an important step towards harnessing the unexplored potential of Danish EHRs as the technology could be used to enhance the accessibility of crucial patient information stored within the narrative text of EHRs. This could help enhance clinical research possibilities and improve patient treatment.

1.1 Objective

The objective of this PhD dissertation is to develop language technology, specifically pre-trained language models, that can be used to classify and extract clinically important information from the narrative text of Danish electronic health records. Moreover, this dissertation aims to strengthen research in Danish clinical language technology by developing language resources that can be used to develop and evaluate language models. Finally, the dissertation explores possible bias issues when developing clinical language models, specifically focusing on how bias can arise due to the distribution of patient subgroups within the dataset used to develop language models.

1.1.1 Research questions

1. How do Danish clinical pre-trained language models perform compared to Danish open-source pre-trained language models and other deep learning-based architectures?
2. How can bias in clinical language models arise from the representation of patient subgroups in the datasets used for model development?
3. How can clinical language models contribute to extracting information from Danish EHRs?

1.2 Contributions

This is an article-based dissertation. During my PhD-studies, I have published three journal papers and six conference papers in collaboration with Danish and international researchers from the fields of medicine and computer science.

The list below shows the publications included in this dissertation.

- Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Cristina Soguero-Ruiz, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, and Pernille Just Vinholt. "Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. © 2022 IEEE. Reprinted, with permission from the authors. 2022, pp. 1–4
- Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Pernille Just Vinholt, and Thiusius Rajeeth Savarimuthu. "MeDa-BERT: A medical Danish pretrained transformer model". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 301–307. URL: https://aclanthology.org/2023.nodalida-1.31

- Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Pernille Just Vinholt, Anne Bryde Alnor, and Thiusius Rajeeth Savarimuthu. "Investigating anatomical bias in clinical machine learning algorithms". In: *Findings of the European Chapter of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023

- Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, Cathrine Brødsgaard Nielsen, Lou-Ann Christensen Andersen, Søren Andreas Just, and Pernille Just Vinholt. "Deep learning detects and visualizes bleeding events in electronic health records". In: *Research and practice in thrombosis and haemostasis* vol. 5, no. 4, 2021, e12505

- Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Rasmus Søgaard Hansen, Thiusius Rajeeth Savarimuthu, Rasmus Bank Lynggaard, and Pernille Just Vinholt. "Doctors identify hemorrhage better during chart review when assisted by artificial intelligence". English. In: *Applied Clinical Informatics, © Georg Thieme Verlag KG*, July 2023. ISSN: 1869-0327. DOI: 10.1055/a-2121-8380

- Martin Sundahl Laursen, Jannik skyttegaard Pedersen, Rasmus Hansen, Thiusius Rajeeth Savarimuthu, and Pernille Just Vinholt. "Danish Clinical Named Entity Recognition and Relation Extraction". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 655–666. URL: https://aclanthology.org/2023.nodalida-1.65

The list below shows the publications published during my PhD-studies that are not part of this dissertation.

- Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Pernille Just Vinholt, Rasmus Søgaard Hansen, and Thiusius Rajeeth Savarimuthu. "Benchmark for Evaluation of Danish Clinical Word Embeddings". In: *Northern European Journal of Language Technology* vol. 9, no. 1, 2023

- Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Søren Andreas Just, Thiusius Rajeeth Savarimuthu, Brian Blomholt, Jakob KH Andersen, and Pernille Just Vinholt. "Factors Facilitating the Acceptance of Diagnostic Robots in Healthcare: A Survey". In: *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2022, pp. 442–448

- Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Pernille Just Vinholt, and Thiusius Rajeeth Savarimuthu. "Automatic Annotation of Training Data for Deep Learning Based De-identification of Narrative Clinical Text". In: *WNLPe-Health 2022: Proceedings of The First Workshop on Context-aware NLP in eHealth (WNLPe-Health 2022)*. CEUR Workshop Proceedings. 2023, pp. 30–44

1.3 Dissertation outline and chapter overview

In summary, the contribution of this dissertation is the development of language technology that can extract clinically important information from the narrative text of Danish EHRs. My contributions include acquiring large EHR datasets that can be used to pre-train language models, creating large annotated datasets to compare and evaluate language models, exploring a specific kind of bias that can arise due to the data curation methodology, and exploring possible use-case scenarios of pre-trained language models in clinical settings.
The dissertation is organized as follows:

Chapter 2 describes the background for understanding the work presented in the dissertation. It provides a broad overview of the transition from deep learning-based text encoders such as the RNN to the introduction of pre-trained language models. Next, it gives an overview of language models that can be used for the clinical domain and possible bias issues. Moreover, it provides a short overview of previous work within Danish clinical language technology and the challenges that involve developing the models for this domain.

Chapter 3 presents work on developing two pre-trained language models that can be used for Danish clinical NLP tasks. The models are evaluated on clinical classification tasks and compared with other language models to address the first research question.

Chapter 4 presents work on a specific type of bias called anatomical bias that can arise for clinical classification tasks due to the data curation methodology. This chapter addresses the second research question.

Chapter 5 addresses the third research question by presenting language models that can be used to extract clinically relevant information from the narrative text of Danish EHRs. First, the chapter presents work on extracting bleeding events. Next, the chapter presents a user study that investigates MDs' performance when using a bleeding identification algorithm in a clinical scenario and the MDs' perception of using it. Finally, the chapter presents a dataset and a language model that can be used to extract clinical events, attributes, and relations between events.

Chapter 6 summarizes the findings of the dissertation and presents possible directions for future research in Danish clinical language technology.

CHAPTER 2

Background and related work

This chapter describes the background for understanding the work presented in the dissertation. The chapter provides a broad overview while detailed descriptions and technical details are presented in later chapters.

Most of my work has been centered around deep learning-based NLP which has developed at a rapid pace throughout my PhD-studies. The rapid pace in NLP has primarily been due to the introduction of a neural network architecture called the transformer [24]. As the transformer is a recent development, part of my work has involved comparing it with older architectures such as RNNs that was previously the preferred architecture for many NLP tasks. Section 2.1 presents a short overview of the transition from traditional deep learning-based architectures to the introduction of the transformer model and pre-trained language models.

As described in the introduction, most innovation in NLP has emerged in the general domain where abundant text is available to train the language models. However, development of clinical NLP quickly followed as many of the techniques can be directly applied to this domain. Section 2.3 gives an overview of clinical language models and how general domain language models can be adapted to the clinical domain.

Although clinical NLP is developing in line with the innovations from the general domain, most of the research focuses on the English language. Small-resource languages like Danish are lagging behind, especially for specialized domains such as the clinical. Section 2.4 provides an overview of previous work within Danish clinical NLP.

Next, Section 2.5 describes the risks of applying language models in clinical settings due to their vulnerability to reflect biases present in the data used to train the models.

Finally, Section 2.6 presents a potential use-case scenario for clinical language models, explaining how they can offer valuable information for supporting bleeding assessments in the clinic and for conducting clinical research.

2.1 Traditional deep learning-based text encoders

From roughly 2012, deep learning-based methods combined with an increasing amount of unlabeled and labeled text data provided significant advantages over former methods. In 2012, Mikolov et al. [25, 26] presented a big breakthrough in NLP with the introduction of Word2Vec. This approach made it possible to represent words as positions in large-dimensional spaces from which word similarity could be measured using distances in the space. Most importantly, the vector representations, known as word embeddings, were created without the use of labeled data. Instead, they were developed using a small neural network that constructs meaningful word embeddings by predicting the center or context word within a small window size in a large-scale corpus. In 2014, Pennington et al.[27] developed another method for creating word embeddings called GloVe. Instead of predicting words within a small window size, the word embeddings were created by predicting probability ratios of global word co-occurrence counts in a large text corpus.

The Word2Vec and GloVe methods were limited in that they could only create word embeddings for words that existed in the pre-training corpus. When applying these embeddings to downstream tasks, the model may struggle to understand the meaning of out-of-vocabulary words. This limits the performance of embedding-based models that deal with rare or domain-specific words. In 2017, Bojanowski et al. [28] developed the FastText model that addressed this limitation. Instead of creating word embeddings for whole words only, the FastText method also trained word embeddings for the most common n-gram characters of all the words in the corpus. For example, for the word *bleeding*, they trained word embeddings for both *ble*, *lee*, *eed*, *edi*, *din*, *ing*, and *bleeding*. Words that were not present in the pre-training corpus could then be created by adding all the individual character n-grams of that word.

After creating word embeddings on a large unlabeled corpus, they can be used as input to neural networks to perform a wide range of NLP tasks such as text classification and named entity recognition (NER). Often, the word embeddings are used as inputs to RNN architectures such as the long short-term memory network (LSTM) [29] and gated recurrent unit (GRU) [30] or to a convolutional neural network (CNN).

A CNN uses a set of learned filters to recognize patterns in the input sequence. The filters slide across the input and create a feature map containing the features that they have learned to be important during training. The size of the filters is a hyperparameter, typically set to between one and ten, that determines the model's ability to capture dependencies between neighbor tokens. After the filters have convolved across the input and created the feature map, neural network layers such as max-pooling and fully-connected layers can be applied to make a sentence representation.

RNNs are another type of neural network architecture designed to model sequential problems. They work by processing the input tokens one element at a time while maintaining a memory of the previous inputs. This memory makes RNNs capable of modeling dependencies between words, which is an important ability in many NLP tasks. After having processed the

input, the final hidden memory can be used as a sentence representation that can be used in tasks like text classification.

2.2 Transformer-based text encoders

In 2018, a new paradigm in NLP occurred with the introduction of the transformer architecture [24]. The transformer was originally developed for machine translation where it achieved SOTA results. The architecture consists of an encoder and decoder block. The encoder block maps the input words to word embeddings which are used by the decoder to output a sequence of words. Unlike previous sequential neural networks, the transformer relied only on an attention mechanism to capture global dependencies between words. This made it possible to effectively model long-term dependencies and to process text more efficiently [31, 32]. Moreover, the attention mechanism made it possible to create *contextual* word embeddings. With contextual word embeddings, each word is not treated as an independent entity with a fixed meaning as seen in the Word2Vec and GloVe methods. Instead, the vector representation of the word changes based on the context in which it appears. This makes it possible for the model to represent word ambiguity and fine-grained nuances of meaning for a word when used in different contexts [33].

Shortly after its introduction, researchers found that transformer architectures benefited greatly from an initial self-supervised pre-training stage followed by a supervised fine-tuning stage [4]. In the pre-training stage, the model learns to represent and infer the meaning of the written language by being exposed to a large amount of text, e.g., by predicting the next word in a sentence or by filling in a masked word. Next, the model can use the learned representations as is or fine-tune them for specific tasks such as information extraction in EHRs. Pre-trained transformer models can be divided into three families depending on their use of the encoder and decoder: encoder-only, decoder-only, and encoder-decoder models [34]. This dissertation focuses specifically on the encoder-only models, also known as autoencoding models, which use only the encoder block of the original transformer architecture.

One of the first pre-trained transformer models was BERT (Bidirectional Encoder Representations from Transformers) [4]. The largest BERT model consisted of 24 transformer layers and was pre-trained on 3.3 billion words from the general domain such as Wikipedia. In the pre-training stage, BERT jointly optimizes a masked language modeling (MLM) and next sentence prediction (NSP) objective. In MLM, a random subset of tokens in the input sequence is replaced with a [MASK] token, and the objective of the model is to predict the masked tokens. In NSP, the model predicts whether two text segments occur next to each other in the original text. BERT achieved SOTA results on 11 NLP tasks, including General Language Understanding Evaluation (GLUE) [35] and Stanford Question Answering Dataset [36].

Liu et al. [37] replicated the BERT study and introduced an improved transformer model called RoBERTa (A Robustly Optimized BERT Pre-training Approach). RoBERTa improved over BERT by pre-training a larger model on more data, for a longer time, and with larger batch sizes. Moreover, they removed the NSP objective as it did not improve the model's performance.

Pre-trained models such as BERT and RoBERTa achieved SOTA results, but they are prohibitively expensive to develop. The BERT model was pre-trained on 64 TPUs for 4 days and RoBERTa was pre-trained on 1024 Nvidia V100 GPUs for one day. This compute budget is beyond reach for most companies and universities. Consequently, much research has focused on reducing the computational requirements for training these pre-trained models while preserving their powerful capabilities [38–41].

One successful approach was the ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) model [38]. This model replaced the commonly used MLM objective with replaced token detection (RTD). In RTD, the model is developed using two sub-models. The first sub-model is trained as a generator that corrupts the input text by replacing words with other plausible words. Next, the other sub-model predicts whether each token in the corrupted input was replaced by the generator or not. The resulting ELECTRA models performed comparably to RoBERTa using only ¼ of the compute. Moreover, the study showed that a small ELECTRA model trained for one day on a single GPU outperformed a comparably small BERT model by five percentage points on the GLUE benchmark.

2.3 Language technology for the clinical domain

As described in the previous section, pre-trained language models achieved SOTA results on many NLP benchmarks and, therefore, they have become the backbone architecture for many modern NLP-powered applications. However, directly applying the pre-trained language models to specialized domains such as the clinical provides unsatisfactory results [42, 43]. Clinical information extraction can be particularly challenging because the language used in EHRs differs significantly from that of the general domain. For example, EHR text contains domain-specific abbreviations, rare medical terminology, and has a high rate of spelling and typing errors [13, 14]. Moreover, the text contains references to individuals other than the patient and there is a frequent use of negations that rejects clinical symptoms and diseases. In order to capture the terminology and style of the specialized domain, clinical language models are often pre-trained on text from EHRs by either training them from scratch or from an existing checkpoint, e.g., a language model pre-trained on the general domain.

Alsentzer et al. [44] developed two English clinical pre-trained language models using a corpus of approximately 2 million clinical notes. The models were initialized from either a pre-trained general domain BERT model or a bio-medical domain BERT model. They found that the fine-tuned clinical models performed better than general domain models on tasks such as clinical text classification and clinical NER. Huang et al. [45] used the same pre-training dataset to develop an English clinical BERT model and found that it performed better than a general domain BERT model on a language modeling task, a clinical similarity task, and for hospital readmission prediction. Similarly, other papers [43, 46, 47] found that English in-domain medical language models perform better than models trained on the general domain.

2.4 Language technology for the Danish clinical domain

For the Danish language, there has been limited research into the performance of both clinical word embeddings and clinical language models [48]. This is partly due to a lack of large clinical text corpora and datasets that can be used for the development and evaluation of the models. Text from EHRs contains sensitive patient information causing legal and ethical considerations that limit the possibility of sharing datasets and models. Therefore, researchers are often forced to collect their own large-scale datasets for pre-training language models and to annotate data that can be used for evaluating the models on downstream tasks such as text classification and NER.

Some research has been conducted in Danish clinical language processing. Eriksson et al. [49] created a pipeline for identifying possible adverse drug events (ADE) from the unstructured narrative text of Danish EHRs. The authors build an ADE dictionary and used rule-based approaches to extract the possible ADEs from the EHR text. Roque et al. [50] used a rule-based approach to extract clinically relevant terms and map them to ICD-10 disease codes. Similarly, other papers have used rule-based approaches for text mining in Danish EHRs [51, 52]. Recently, machine learning approaches have also been investigated. For example, Enevoldsen et al. [53] used machine learning approaches such as XGBoost [54] and Naive Bayes to classify if symptoms described in Danish clinical notes were COVID-19 pandemic-related or not. Moreover, Kaas-Hansen et al. [55] used FastText word embeddings and small neural networks to predict associations between free-text information and medication exposure. However, there have been limited efforts in creating and evaluating pre-trained transformer language models developed on a large-scale corpus of Danish EHRs.

2.5 Bias in clinical language models

While clinical language models could help advance clinical research and practice, they also hold the risk of exacerbating bias which is systematic and unfair discrimination against certain individuals or groups of individuals in favor of others [56].

During pre-training, language models learn to represent the meaning of words and phrases through repeated exposure to a vast amount of text data. While this approach has shown to be very effective for representing the meaning of language, it also carries a potential risk of injecting biases from the text corpus into the model's representations of language. Similarly, if biased data is used to fine-tune a model for a specific task, it might also be reflected in the model's predictions [10]. To prevent this, much research has been dedicated to uncovering and mitigating bias in pre-trained language models. For example, Oberheimer et al. [57] observed bias in an algorithm used to guide health decisions in the U.S. healthcare system. The authors found evidence that the algorithm falsely predicted that black patients were healthier than equally sick white patients and estimated that this racial bias reduced the number of black patients identified for extra care by more than half. Similarly, other studies have found algorithms that

contained gender and racial biases [58–60]. Therefore, language models must be developed using carefully curated datasets to avoid encoding biases into the models as this could have fatal consequences in clinical settings [10].

2.6 Language models for extraction of bleeding events

One area in which clinical language models could add value is for extracting regarding bleeding events from the narrative text of EHRs. Bleeding is a life-threatening condition that occurs within 14 days of admission for 3.2% of medical patients. It is, therefore, recommended to assess bleeding risk in all patients admitted to the hospital [61]. Information about past bleeding events and the specific location of bleeding can be used to predict the risk of future bleeding and to guide anticoagulant treatment. However, obtaining this information is prohibitively time-consuming because ICD-10 codes for bleeding are unreliable and therefore MDs are required to manually look through the narrative text of the EHR [62]. Consequently, previous studies [63–65] have shown a lack of adherence to risk scores that include previous bleeding events as a factor, mainly because of limited time resources, especially in acute clinical settings where direct patient care is prioritized. Automatic extraction of previous bleeding events would facilitate a fast way of assessing bleeding risk on an informed basis, reduce the amount of time medical professionals spend manually reviewing EHRs, and improve patient safety.

Automatic extraction of bleeding events in the narrative text of EHRs is a complex task. Bleeding events can be described in numerous ways and the description and choice of bleeding-indicating words often depend on the location of the bleeding. Moreover, because of the unstructured and noisy nature of the EHRs, bleeding-indicating phrases can include misspellings, grammatical ambiguity, and MD-specific term abbreviations. In addition, bleeding events can be negated in multiple ways, creating another layer of complexity.

Extraction of bleeding events has been investigated for the English language. For example, Li et al. [66] used a CNN and an LSTM to extract sentences describing bleeding in English EHRs with an F1 score of 94%. Taggert et. al [67] used a rule-based approach to extract medical notes describing bleeding events in English EHRs with an F1 score of 74%. Mitra et al. [68] used a BERT-based model to extract token entities that indicate bleeding with an F1 score of 75%. In a Danish context, a limitation of the aforementioned studies is that the algorithms were developed using the English language and thus cannot be applied for Danish EHRs. Additionally, the studies used a relatively small test set to evaluate their models (<1000 samples), and they did not investigate the performance of the models across different bleeding types or patient characteristics.

CHAPTER 3

Development of Danish pre-trained language models for the clinical domain

This chapter presents two papers describing the development and evaluation of a Danish clinical language and a Danish medical language model. I use the term *clinical* to denote a language model that has been pre-trained on EHRs and *medical* for language models that have been pre-trained on medical literature such as medical books or medical text from the internet.

First, Section 3.1 describes the transformer architecture in detail as it is used as the backbone architecture for the developed models.

Next, Section 3.2 presents the paper *Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records* [15]. For this paper, we developed a Danish clinical pre-trained language model called Clin-ELECTRA and compared its performance with a BERT model trained on the general domain and an LSTM model.

Section 3.3 presents our paper *MeDa-BERT: A medical Danish pretrained transformer model* [16]. In this paper, we collected a large medical corpus and used it to develop a medical Danish language model (MeDa-BERT) that could be shared openly with the NLP community.

3.1 The transformer model

The two developed models, Clin-ELECTRA and MeDa-BERT, consist of transformer layers [24]. A transformer layer transforms input tokens into contextualized word representations that can be used to make word and sentence classifications. Figure 3.1 illustrates the transformer architecture.

First, input tokens are transformed into word embeddings and added with positional embeddings. Positional embeddings are added to inject information about the position of the

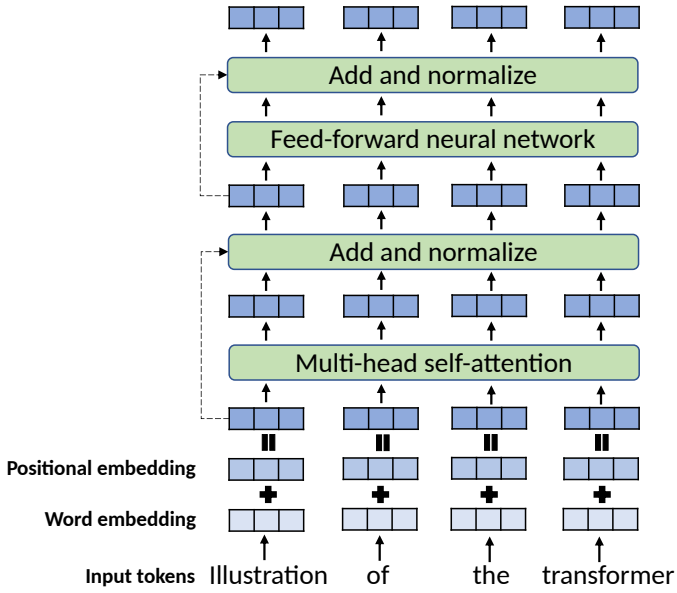


FIGURE 3.1: Illustration of the transformer architecture.

tokens in the input sequence. Without positional embeddings, the input tokens to the transformer would be permutation invariant and the model would thereby perceive the tokens as an unordered sequence. In the original transformer paper, the positional embeddings were hardcoded using sine and cosine functions of different frequencies. In models such as BERT and ELECTRA, the positional encodings are initialized at random and learned during training, which produces similar results as the hardcoded approach [24].

Next, the word representations are used as input to a self-attention layer. In self-attention, tokens can inject information from other tokens into their representation, thereby creating contextualized word embeddings. To do this, the self-attention layer creates three representations from the input word embedding: a query, key, and value. These representations are used to calculate the output of the embedding, which is a single contextualized word embedding computed as a weighted sum of the computed values. The weighted values are computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are matrices containing the query, key, and value vectors for all input tokens, and d_k is the number of dimensions for the keys.

A single-head attention layer is illustrated in Figure 3.2, where the contextualized word embedding for the first token in the input sequence is calculated.

To expand the transformer's capability of representing a token in the context it appears in, the transformer uses multi-head attention. With multi-head attention, the model creates

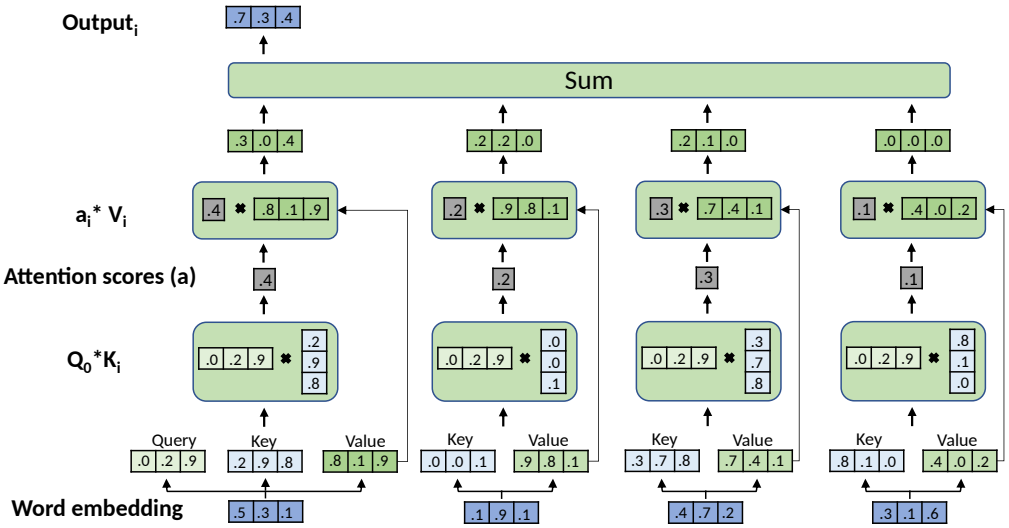


FIGURE 3.2: Illustration of the single-head attention mechanism. The contextualized word embedding for the first word is created by creating a query, key, and value vector. The query vector is multiplied by the key vectors from all the other words which results in four attention scores. The attention scores are multiplied by their respective value vector, creating the *weighted* values. Next, the weighted values are added to provide the final contextualized word embedding for the first word in the sequence. The same approach is utilized for creating the contextualized word embeddings of the remaining input words. For simplicity, the figure does not show the $\frac{1}{\sqrt{d_k}}$ and $softmax$ function which is used to calculate the attention scores.

different views, referred to as heads, of the same tokens by attending to different positions in its context. To do this, multiple sets of query, key, and value vectors are created for each token, using weight matrices W^Q, W^K, and W^V, respectively. Next, the resulting vector from each head is concatenated and projected by a weight matrix, W^O, to give the final output of the layer:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{3.1}$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.2}$$

The self-attention layer is followed by a residual connection [69] and a normalization layer [70] calculated as:

$$\text{output} = \text{normalize}(x + \text{Sublayer}(x)) \tag{3.3}$$

where x is the input embeddings to the multi-head self-attention layer and $Sublayer(x)$ is the output of the self-attention layer.

Next, the embeddings are used as input to a small feedforward neural network (FNN) with a ReLU activation function:

$$\text{FNN(x)} = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3.4}$$

Finally, the output of the FNN is added to the input of that layer and normalized as in eq. (3.3).

The steps described above constitute a transformer layer. Typically, multiple layers are stacked when developing pre-trained language models. For example, the models presented in the sections below contain 12 transformer layers.

3.2 Clin-ELECTRA: A language model pre-trained on Danish Electronic health records

This section presents our paper *Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records* [15]. The objective of this paper was to compare clinical language models with an open-source BERT model trained on text from the general domain. The paper focuses mainly on comparing the performance of the models on a bleeding site prediction task and does not describe the development of the pre-trained Clin-ELECTRA model in detail. Therefore, this section first describes how the model was pre-trained and fine-tuned after which the paper is presented.

3.2.1 Development of Clin-ELECTRA

As described in section 2.2, the ELECTRA pre-training approach is an efficient method for developing language models which benefit research teams with limited compute resources [38]. Since we had a limited compute budget, we developed an ELECTRA model with 13.5 million parameters that could be pre-trained in five days on a single Nvidia V100 GPU. Although Clin-ELECTRA is significantly smaller than other pre-trained language models, we have found it to work well for many clinical NLP tasks.

The development of Clin-ELECTRA consisted of four steps, 1) data collection, 2) development of a custom tokenizer, 3) Sample creation, and 4) pre-training. Each step is described below.

Data collection

In order to pre-train Clin-ELECTRA we acquired a large clinical text corpus consisting of narrative text from 299,718 Danish EHRs from Odense University Hospital (OUH). The corpus contained text from all departments of OUH and included a total of 1,411,575,062 tokens.

Custom tokenizer

A tokenizer is used to break down a text corpus into its individual units, often referred to as tokens. For example, these tokens can be words, sub-words, numbers, or non-alphanumeric characters such as commas and question marks. Tokenization is a crucial initial step as the tokens generated by the tokenizer serve as the input to the language model that transforms each token into a word embedding. Creating distinct word embeddings for the words in a large corpus would take up a prohibitively large amount of memory. Therefore, language models are often developed to represent the most frequent words and sub-words of a text corpus. Words that are in the language model's vocabulary are represented by a specific word embedding for that word while out-of-vocabulary words are split into the sub-words that are present in the language model's vocabulary. Figure 3.3 visualizes this approach. In the illustration, the vocabulary of the model does not contain the word *unknown*. Therefore, the tokenizer splits the word into two sub-words, *un*## and ##*known*, which are in the model's vocabulary.

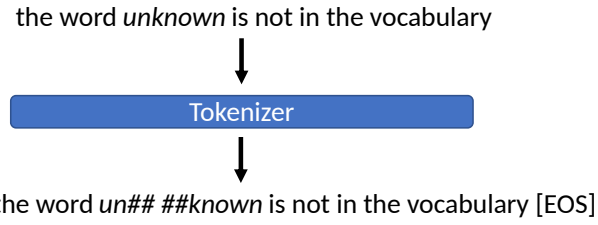


FIGURE 3.3: The tokenizer splits the word unknown into its sub-words and adds the special start-of-sentence [SOS] and end-of-sentence [EOS] tokens. ## denotes a sub-word.

For Clin-ELECTRA, we created a vocabulary of 30,522 words and sub-words using the WordPiece method [71, 72]. WordPiece constructs a vocabulary from a text corpus by splitting all the initial words into their individual sub-words. For example, [“word”] is split into [“w”, “##o”, “##r”, “##d”], where ## indicates a sub-word. Next, WordPiece iteratively merges the word pair with the highest score according to:

$$\text{score} = \frac{\text{freq\_of\_pair}}{\text{freq\_of\_first\_element} \times \text{freq\_of\_second\_element}} \tag{3.5}$$

WordPiece thereby prioritizes merging word pairs that are more frequent when combined than when separated. For example, WordPiece will not necessarily merge [“un”, “##able”] if the individual sub-words occur frequently in the corpus. On the other hand, one could imagine that ["pa", "##tient"] would be merged to ["patient"] since that word would be relatively frequent in an EHR corpus compared to the two individual sub-words. Once a desired vocabulary size is acquired, the tokenizer uses the vocabulary to tokenize a text corpus. Out-of-vocabulary words are split using the longest sub-word in the tokenizer's vocabulary. If a token cannot be represented, it will be tokenized as the special token [UNK].

We used the WordPiece algorithm on the collected EHR corpus from OUH and created a vocabulary of 30,522 tokens of which 7796 were sub-words.

Sample creation

To pre-train an ELECTRA model, text from the EHRs must be separated into individual samples. Our samples were created by merging sentences from the same headline of an EHR note up to 128 tokens, which is the maximum input length for Clin-ELECTRA. Next, each sample was prepended with a [SOS] token, and a [EOS] token was added at the end of the sample, as seen in Figure 3.3. The [SOS] token indicates the start-of-sentence and the [EOS] token indicates the end-of-sentence to the model.

Finally, samples within the same EHR note with a length of less than 128 tokens were merged. Once all the samples were processed by the tokenizer, tokens were represented as their word embedding vector with 128 dimensions.

Pre-training Clin-ELECTRA

We pre-trained Clin-ELECTRA using replaced token detection, visualized in Figure 3.4.

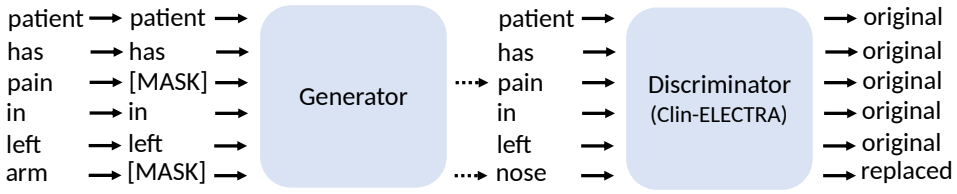


FIGURE 3.4: Visualization of the ELECTRA pre-training approach. The generator predicts the masked tokens and the discriminator predicts if the tokens in its input have been replaced by the generator or not. The illustration is created with inspiration from [38]

.

During pre-training, two sub-models are used to create an ELECTRA model: a generator and a discriminator. Both sub-models consist of 12 transformer layers, but the generator only has ¼ of the parameters of the discriminator.

The generator receives a tokenized input sample with some of the words replaced with a special [MASK] token. The objective of the generator is to predict the correct token for each [MASK] token, i.e., the generator is trained with MLM. The output of the generator is used as input to the discriminator that predicts whether each token in the input was replaced by the generator or not. For example, in Figure 3.4, the discriminator should predict that the word *nose* is not from the original sentence. The loss of the generator and discriminator is added and used to iteratively update the parameters of both sub-models during pre-training.

We pre-trained Clin-ELECTRA with a batch size of 128 for 1.5 million steps corresponding to 17 epochs on the EHR corpus, i.e., the model was exposed to 24.5 billion tokens during pre-training. All pre-training and model parameters can be seen in Table 3.1.

Parameter	Value
Architecture	
Number of layers	12
Hidden size	256
FNN inner hidden size	1024
Attention heads	4
Attention head size	64
Embedding Size	128
Generator size (multiplier for hidden-size, FNN-Size, and num-attention-heads)	1/4
Max seq. length	128
Optimization	
Learning rate	1e-4
Optimizer	Adam
Adam epsilon	1e-6
Adam beta1	0.90
Adam beta2	0.999
Dropout	0.1
Attention dropout	0.1
Learning rate decay	Linear
Batch size	128
Warm up	1000
Epochs	17
Train steps	1.5M
Mask percentage	15

TABLE 3.1: Architecture and optimization parameters for pre-training Clin-ELECTRA

After pre-training, only the discriminator is used for downstream tasks. For sentence prediction tasks such as bleeding site classification, the model uses the [SOS] token as a sentence representation and makes a prediction using a classification layer, e.g., a single fully-connected layer as seen in Figure 3.5. Then, during fine-tuning, the model iteratively updates its parameters to create a meaningful sentence representation that can be used for prediction tasks.

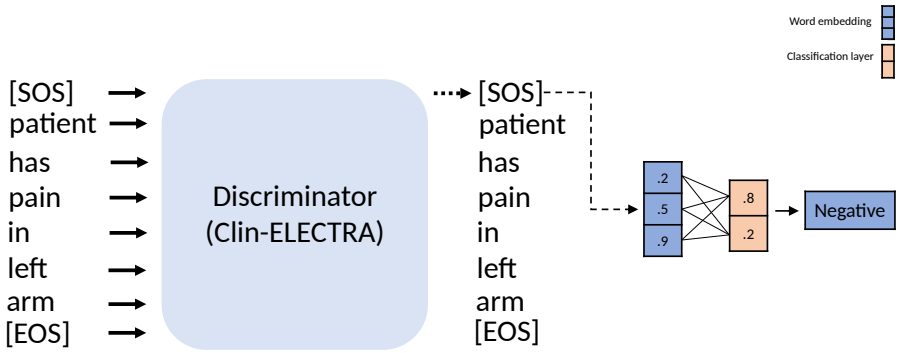


FIGURE 3.5: Example of Clin-ELECTRA used for a bleeding prediction task. The model takes the sentence as input and uses the word embedding of the [SOS] token as a sentence representation. The sentence representation is followed by a fully-connected classification layer and predicted as negative for bleeding.

In our paper *Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records* [15], we used the developed Clin-ELECTRA to

classify the anatomical location of bleeding events in EHR paragraphs. From a medical perspective, being able to classify the anatomical location is important in bleeding assessment. The bleeding site is important because it is used as a parameter in distinguishing between major and minor bleeding events in risk score calculations [61], where major bleeding is associated with an elevated risk of future bleeding. Therefore, we constructed a dataset consisting of paragraphs describing bleedings and labeled it with the bleeding site.

To evaluate the performance of Clin-ELECTRA, we compared its results with a general domain BERT model and an LSTM model. We found that Clin-ELECTRA performed better than the BERT and LSTM model for bleeding site classification. Thereby, the results show the advantage of developing in-domain clinical language models compared to using open-source models, even when only limited compute budgets are available to pre-train the in-domain models.

Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records

1st Jannik S. Pedersen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
jasp@mmmi.sdu.dk

2nd Martin S. Laursen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
msla@mmmi.sdu.dk

3rd Cristina Soguero-Ruiz
Signal Theory and Communication
Universidad Rey Juan Carlos,
Fuenlabrada, Spain
cristina.soguero@urjc.es

4th Thiusius R. Savarimuthu
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
trs@mmmi.sdu.dk

5th Rasmus Søgaard Hansen
Department of Clinical Biochemistry
Odense University Hospital
Odense, Denmark
Rasmus.Sogaard.Hansen@rsyd.dk

6th Pernille J. Vinholt
Department of Clinical Biochemistry
Odense University Hospital
Odense, Denmark
Pernille.Vinholt@rsyd.dk

2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-8791-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BHI56158.2022.9926955

***Abstract*—Bleeding can be a life-threatening condition which occurs for 3.2% of medical patients. Information about previous bleeding and bleeding site is used to predict the risk of future bleeding and guide anticoagulant treatment. However, obtaining this information is a time-consuming task as it is contained in the free text of electronic health records.**

Previous research has mainly been focused on extracting bleeding events but does not classify the bleeding site which is important for assessing the severity of the bleeding. This study creates the first dataset for developing and evaluating machine learning models for classification of bleeding site. The dataset consists of sentences annotated by medical doctors as belonging to one of ten bleeding sites. The sentences were annotated in 149,523 electronic health record notes from 1,533 patients of Odense University Hospital, Denmark, between 2015 and 2020.

We compare different deep learning models on classifying bleeding site and find that a ∼13M parameter ELECTRA model pretrained on clinical text achieves higher accuracy (0.905 ± 0.002) than a ∼110M parameter general BERT model (0.884 ± 0.001) on a balanced test set of 1,500 sentences.

We furthermore test different methods for dealing with unbalanced data without finding any significant differences between methods.

***Index Terms*—Electronic health records, natural language processing, deep learning, transformer, BERT**

I. Introduction

Bleeding is a dangerous event causing increased morbidity and mortality [1], [2]. Within 14 days of admission, 3.2% of medical patients experience in-hospital bleeding, and approximately one third are considered major bleedings [3] which can be fatal.

Previous clinically relevant bleeding is a strong independent predictor for future bleeding. Bleeding history is part of the IMPROVE score [3] for determining bleeding risk in admitted medical patients and HAS-BLED [4] for addressing bleeding risk before prescription of anticoagulation therapy. The IMPROVE score defines major bleedings as those occurring within a critical organ such as intracranial, adrenal gland or spinal bleeding. Thus, information about bleeding site is an important factor for patient treatment.

Manually obtaining information about previous bleeding and its site is prohibitively time-consuming since medical doctors (MD) have to manually look through the free text of electronic health records (EHR). Automating this task would facilitate a fast way of assessing bleeding risk on an informed basis as well as decrease the time taken up by MDs manually reading through EHRs.

Several studies [5]–[7] have shown that bleeding events can be extracted from EHRs with high accuracy using deep learning techniques such as Long Short-Term Memory [8] (LSTM) networks, Convolutional Neural Networks [9] and transformer [10] variants. However, these studies do not classify the site of bleeding events. One study [11] does extract information about bleeding site using named entity recognition but does not classify the bleeding site into distinct categories.

In this paper, we create a bleeding site classification dataset using the free text of Danish EHRs. We use this dataset to train and compare an LSTM model and two transformer-based models, BERT [12] and ELECTRA [13]. We find that an in-domain ELECTRA model with ∼13M parameters predicts bleeding site with an accuracy of 0.905 which is significantly better than a general domain BERT model with ∼110M parameters. Moreover, we find no significant performance difference between a ∼9M parameter LSTM model trained with clinical word embeddings and the BERT model. The developed models could potentially be integrated with previously developed bleeding mention detection models [5]–[7] to enrich

*Equal contribution

TABLE I
DISTRIBUTION OF THE BLEEDING SITE CLASSIFICATION DATASET.

	Train	Validation	Test	Sum
Airways	130	150	150	430
Cerebral	1,634	150	150	1,934
Ear-nose-throat	1,161	150	150	1,461
Eyes	971	150	150	1,271
Gastrointestinal	2,771	150	150	3,071
Gynecological	953	150	150	1,253
Internal bleeding	976	150	150	1,276
Skin	1,044	150	150	1,344
Urogenital	2,691	150	150	2,991
Unknown	2,003	150	150	2,303
Sum	14,344	1,500	1,500	17,344

information about bleeding risks in clinical decision support systems.

II. METHODS AND MATERIAL

A. Dataset

The dataset was created using a database of EHRs from Odense University Hospital, Denmark, between 2015 and 2020. The extracted EHRs were distributed between 13 MDs who annotated all sentences as either positive or negative for bleeding. A total of 149,523 EHR notes from 1,533 patients were annotated. This process resulted in 17,344 unique bleeding-positive sentences. Next, all the positive sentences were annotated with a consensus label from three MDs and grouped into one of ten bleeding categories: airways, cerebral, ear-nose-throat, eyes, gastrointestinal, gynecological, internal, skin, urogenital, and unknown. A preprocessing stage was done to remove superfluous spaces and special characters. After preprocessing, the sentences had an average length of 12.3 tokens.

In order to build the deep learning models, the dataset was split into train, validation, and test sets. The validation and test sets include an equal amount of sentences (n=150) from each bleeding site. The remaining were used to train the models. The dataset distribution can be seen in Table I.

B. Deep learning models

We train a LSTM model consisting of a bidirectional LSTM layer with a hidden layer size of 512. The last hidden state of the LSTM is followed by a dropout layer with probability 0.2, a dense layer of size 256, a ReLU activation function, a dropout layer of probability 0.2, and a dense classification layer. The input to the LSTM model is 300-dimensional FastText [14] word embeddings pretrained on EHRs consisting of 1.4B tokens.

Moreover, we fine-tune two transformer models, Danish BERT[1] and a locally available clinical ELECTRA model (Clin-ELECTRA). The BERT model has ∼110M parameters and is pretrained on 1.6B tokens from the Danish general text domain such as CommonCrawl, Wikipedia, and Opensubtitles.

[1]Available from https://github.com/certainlyio/nordic_bert

TABLE II
SUMMARY OF DEEP LEARNING MODELS. WE = WORD EMBEDDINGS.

	Parameters	Pretrain domain	Pretrain size
LSTM	9,484,918	Clinical (WE)	1.4B tokens (WE)
BERT	110,625,034	General	1.6B tokens
Clin-ELECTRA	13,551,370	Clinical	1.4B tokens

The Clin-ELECTRA model has ∼13M parameters and is pretrained on 1.4B tokens from 299,718 Danish EHRs acquired from Odense University Hospital.

We initialize the BERT and Clin-ELECTRA models from their pretrained checkpoints and follow the HuggingFace [15] implementations for text classification. Table II summarizes the deep learning models used in this study.

C. Dealing with unbalanced data

As seen in Table I, the dataset follows a natural unbalance in the amount of sentences from each bleeding site since some bleedings naturally occur more often than others. We compare three methods for training the deep learning models on the unbalanced dataset: unbalanced, upsampling, and training with a weighted loss function.

With the unbalanced method, we simply train each model with the unbalanced training distribution. With upsampling, we upsample all underrepresented classes to the size of the majority class (gastrointestinal), i.e. all categories in the training set are upsampled to include 2,771 sentences. With a weighted loss function, we assign a unique weight to each class when calculating the loss during training. The weight decreases the importance of sentences from classes with many sentences and increases the importance of underrepresented classes. Using the default in Scikit-learn [16], we calculate the weight for each class as

$$w_{c_i} = \frac{n_s}{n_c \cdot n_{s_{c_i}}} \tag{1}$$

where w_{c_i} is the weight for class i, n_s is the total number of sentences, n_c is the number of classes and $n_{s_{c_i}}$ is the number of sentences for class i. We apply w_{c_i} to the categorical cross entropy loss as

$$L = -\sum_{i=1}^{n_c} w_{c_i} \cdot y_i \cdot \log(p_i) \tag{2}$$

where L is the loss function, y_i is the true probability for class i, and p_i is the predicted probability for class i.

D. Model evaluation

We train each model multiple times with different learning rates to compare models based on confidence intervals (CI) and to report statistically significant results. Specifically, for each model we:

1) Train the model with four different learning rates and random initializations.
2) Compute the test set accuracy of the best performing model based on the loss on the validation set.
3) Repeat 1) and 2) five times.

TABLE III
ACCURACY OF THE DEEP LEARNING MODELS ON THE TEST SET USING DIFFERENT METHODS FOR TRAINING WITH UNBALANCED DATA. SE = STANDARD ERROR. CI = 95% CONFIDENCE INTERVAL. UB = UNBALANCED. US = UPSAMPLING. WL = WEIGHTED LOSS.

Model	Mean accuracy ± SE	CI
LSTM-UB	0.865 ± 0.002	0.861-0.869
LSTM-US	0.876 ± 0.002	0.871-0.881
LSTM-WL	0.881 ± 0.001	0.879-0.884
BERT-UB	0.884 ± 0.002	0.881-0.887
BERT-US	0.873 ± 0.003	0.868-0.879
BERT-WL	0.880 ± 0.001	0.878-0.882
Clin-ELECTRA-UB	0.900 ± 0.002	0.896-0.905
Clin-ELECTRA-US	0.903 ± 0.002	0.899-0.906
Clin-ELECTRA-WL	**0.905 ± 0.002**	0.901-0.910

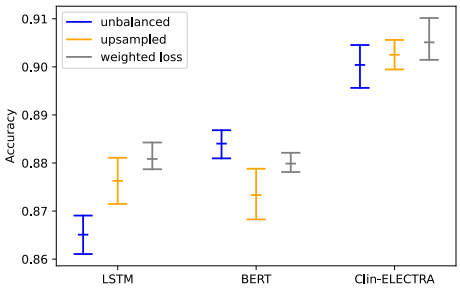


Fig. 1. Mean accuracies and 95% confidence intervals for LSTM, BERT, and Clin-ELECTRA models with the unbalanced, upsampling, and weighted loss methods.

We use the five accuracies to perform bootstrapping with 9,999 replicates to calculate mean accuracy, standard error (SE), and 95% CI for each model. Moreover, we compute the bootstrapped difference of means for each combination of models for evaluation of statistically significant differences in accuracy. For the BERT model, we search for the best model using learning rates 1e-4, 5e-5, 3e-5, and 2e-5. For Clin-ELECTRA and the LSTM model, we use learning rates 4e-4, 3e-4, 2e-4, and 1e-4. BERT and ELECTRA were trained for a maximum of 10 epochs using early stopping. The LSTM model was trained for a maximum of 50 epochs using early stopping.

III. RESULTS

Table III shows the mean accuracy, SE, and CI on the test set for all models and methods for dealing with an unbalanced dataset. Fig. 1 shows the plotted means and CIs. Table IV shows the bootstrapped CIs for difference of means for all model comparisons.

We find that Clin-ELECTRA performs significantly better than the LSTM and BERT models for both the unbalanced (0.900 ± 0.002), upsampling (0.903 ± 0.002) and weighted loss (0.905 ± 0.002) methods. We find no significant difference between the best performing BERT model (0.884 ± 0.002) and LSTM (0.881 ± 0.001).

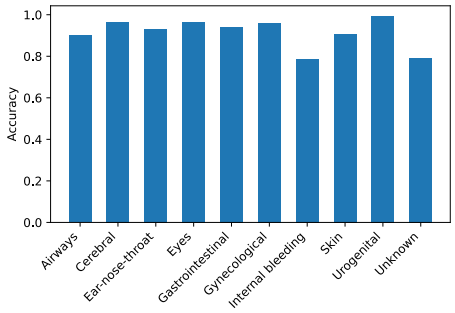


Fig. 2. Per-class accuracy for the best performing Clin-ELECTRA model (weighted loss).

We do not find any of the methods for dealing with unbalanced data (unbalanced, upsampling, or weighted loss function) to perform significantly better when considering results across the three deep learning models. However, for the LSTM model, the upsampling and weighted loss methods perform significantly better than training with the unbalanced dataset. On the contrary, for BERT, the unbalanced and weighted loss methods perform significantly better than the upsampling method. For Clin-ELECTRA, we find no significant performance improvements using any of the methods.

Fig. 2 shows the per-class performance for the best performing Clin-ELECTRA model (weighted loss). It is seen that the model has the worst performance on the internal and unknown bleeding site with accuracies below 0.8. Even though only 130 airways samples are present in the training set, the model performs relatively well with an accuracy of ∼90%.

IV. DISCUSSION

We found that for bleeding site classification, the ∼13M parameter ELECTRA model pretrained on clinical data performs significantly better than the ∼110M parameter BERT model pretrained on general data and the ∼9M parameter LSTM model with clinical word embeddings. Moreover, we found no significant difference between the best BERT and LSTM model.

This result is intriguing since the computational power required for running ELECTRA and the LSTM model is much less than the BERT model. It shows that while increasing the amount of parameters has been shown to improve the performance of deep learning models [17], performance gains can also be reached by focusing on the pretraining data. Due to their pretraining domain, we hypothesize that Clin-ELECTRA and the LSTM model are able to better capture the idiosyncratic language of clinical text which differs from that of the general domain used for pretraining BERT. Clinical notes e.g. include many spelling errors, domain specific

TABLE IV
BOOTSTRAPPED 95% CONFIDENCE INTERVALS FOR DIFFERENCE OF MEANS FOR ALL MODEL COMPARISONS. MEANS ARE COMPUTED AS ROW MINUS COLUMN. * DENOTES A SIGNIFICANT DIFFERENCE AT THE 0.05 LEVEL.

	LSTM-UB	LSTM-US	LSTM-WL	BERT-UB	BERT-US	BERT-WL	ELECTRA-UB	ELECTRA-US	ELECTRA-WL
LSTM-UB	-	(-0.017, -0.005)*	(-0.021, -0.011)*	(-0.024, -0.014)*	(-0.015, -0.002)*	(-0.019, -0.010)*	(-0.041, -0.029)*	(-0.043, -0.032)*	(-0.046, -0.034)*
LSTM-US	(0.005, 0.017)*	-	(-0.010, 0.001)	(-0.013, -0.002)*	(-0.004, 0.010)	(-0.008, 0.002)	(-0.030, -0.018)*	(-0.032, -0.020)*	(-0.035, -0.022)*
LSTM-WL	(0.011, 0.021)*	(-0.001, 0.010)	-	(-0.007, 0.001)	(0.001, 0.013)*	(-0.002, 0.005)	(-0.025, -0.014)*	(-0.026, -0.017)*	(-0.030, -0.019)*
BERT-UB	(0.014, 0.024)*	(0.002, 0.013)*	(-0.001, 0.007)	-	(0.004, 0.017)*	(0.000, 0.008)	(-0.022, -0.011)*	(-0.023, -0.014)*	(-0.027, -0.016)*
BERT-US	(0.002, 0.015)*	(-0.010, 0.004)	(-0.013, -0.001)*	(-0.017, -0.004)*	-	(-0.012, -0.001)*	(-0.034, -0.020)*	(-0.035, -0.023)*	(-0.039, -0.025)*
BERT-WL	(0.010, 0.019)*	(-0.002, 0.008)	(-0.005, 0.002)	(-0.008, 0.000)	(0.001, 0.012)*	-	(-0.025, -0.016)*	(-0.026, -0.019)*	(-0.031, -0.021)*
ELECTRA-UB	(0.029, 0.041)*	(0.018, 0.030)*	(0.014, 0.025)*	(0.011, 0.022)*	(0.020, 0.034)*	(0.016, 0.025)*	-	(-0.008, 0.003)	(-0.011, 0.001)
ELECTRA-US	(0.032, 0.043)*	(0.020, 0.032)*	(0.017, 0.026)*	(0.014, 0.023)*	(0.023, 0.035)*	(0.019, 0.026)*	(-0.003, 0.008)	-	(-0.008, 0.003)
ELECTRA-WL	(0.034, 0.046)*	(0.022, 0.035)*	(0.019, 0.030)*	(0.016, 0.027)*	(0.025, 0.039)*	(0.021, 0.031)*	(-0.001, 0.011)	(-0.003, 0.008)	-

abbreviations, mix of languages (Latin, English, and Danish), and the semantic meaning of words can differ from the general domain [18].

Limitations

While this paper finds that a ∼13M clinical ELECTRA model performs better than a ∼110M parameter BERT model, larger general domain transformer models pretrained on more data might perform better. At the moment, no such model exists for Danish. Applying increasingly larger transformer models, however, also increases the need for expensive computational infrastructure which decreases the usability for researchers with limited access to computational power.

The clinical dataset used to pretrain the ELECTRA model could be used to further pretrain the BERT model which might improve its representations of clinical text.

The classification dataset and clinical ELECTRA model presented in this paper cannot be shared publicly due to privacy concerns but we advise interested researchers to contact us for sharing possibilities.

V. CONCLUSION

This paper created the first dataset to train machine learning models for bleeding site classification in the free text of EHRs. We find that a ∼13M parameter clinical ELECTRA model can classify bleeding site with an accuracy of 0.905, surpassing both a LSTM model and a BERT model which is eight times larger in terms of parameters. Moreover, we compared different methods for dealing with unbalanced datasets. For the methods considered in this paper, we found none of them to work consistently better across all classifiers, suggesting that several approaches should be considered when working with different deep learning architectures with unbalanced data.

REFERENCES

[1] Berger JS, Bhatt DL, Steg PG, Steinhubl SR, Montalescot G, Shao M, Hacke W, Fox KA, Berger PB, Topol EJ, Lincoff AM. Bleeding, mortality, and antiplatelet therapy: results from the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial. American heart journal. 2011 Jul 1;162(1):98-105.

[2] Cook DJ, Griffith LE, Walter SD, Guyatt GH, Meade MO, Heyland DK, Kirby A, Tryba M. The attributable mortality and length of intensive care unit stay of clinically important gastrointestinal bleeding in critically ill patients. Critical care. 2001 Dec;5(6):1-8.

[3] Decousus H, Tapson VF, Bergmann JF, Chong BH, Froehlich JB, Kakkar AK, Merli GJ, Monreal M, Nakamura M, Pavanello R, Pini M. Factors at admission associated with bleeding risk in medical patients: findings from the IMPROVE investigators. Chest. 2011 Jan 1;139(1):69-79.

[4] Pisters R, Lane DA, Nieuwlaat R, De Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest. 2010 Nov 1;138(5):1093-100.

[5] Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, Ferraro J, Bucher BT, Lloyd-Jones DM, Rondina MT, Shah RU. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. JAMA network open. 2018 Oct 5;1(6):e183451-.

[6] Li R, Hu B, Liu F, Liu W, Cunningham F, McManus DD, Yu H. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR medical informatics. 2019 Feb 8;7(1):e10788.

[7] Pedersen JS, Laursen MS, Rajeeth Savarimuthu T, Hansen RS, Alnor AB, Bjerre KV, Kjær IM, Gils C, Thorsen AS, Andersen ES, Nielsen CB. Deep learning detects and visualizes bleeding events in electronic health records. Research and Practice in Thrombosis and Haemostasis. 2021 May;5(4):e12505.

[8] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

[9] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11);2278-2324.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[11] Mitra A, Rawat BPS, McManus D, Kapoor A, Yu H. Bleeding Entity Recognition in Electronic Health Records: A Comprehensive Analysis of End-to-End Systems. AMIA ... Annual Symposium proceedings. AMIA Symposium. 2020 ;2020:860-869. PMID: 33936461; PMCID: PMC8075442.

[12] Devlin J, Chang MW, Lee K, Toutanova K. 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) 2019 Jun (pp. 4171-4186).

[13] Clark K, Luong MT, Le QV, Manning CD. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. International Conference on Learning Representations 2019 Sep 25.

[14] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017 Dec 1;5:135-46.

[15] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771. 2019 Oct 9.

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

[17] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, Schuh P. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311. 2022 Apr 5.

[18] Leaman, Robert, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. Journal of Biomedical Informatics, 57:28–37.

3.3 MeDa-BERT: An Open-domain Danish medical language model

Clin-ELECTRA was pre-trained on text from EHRs and, therefore, the model could not be shared openly with the NLP community. However, sharing models and datasets is important as it can foster innovation and collaborations between researchers who work within the same subfields of NLP. For example, by sharing domain specific language models, researchers can build upon each other's work which could lead to faster and more efficient development and evaluation of language models.

While text from the clinical domain contains sensitive information, text from the medical domain does not. The medical domain provides knowledge about various aspects of the medical field including anatomy, physiology, diseases, symptoms, medications, and medical practices. Although the clinical domain focuses more on the practical aspects of patient care and decision-making, the medical and clinical fields are similar in that they both deal with medical terminology and procedures. For the English language, language models trained on the medical domain have been utilized as base models for continued pre-training on clinical text or directly for clinical tasks with promising results [43, 44, 46]. Developing medical language models for the Danish language would therefore be a valuable asset for both the medical and clinical domains.

In our paper *MeDa-BERT: A medical Danish pretrained transformer model* [16], we collected a Danish medical text corpus and used it to develop a medical Danish BERT model (MeDa-BERT). In contrast to Clin-ELECTRA which was pre-trained with randomly initialized weights, MeDa-BERT was initialized with the weights of an existing Danish BERT model, whereafter it was pre-trained again. We did not train the model from scratch because we deemed the medical corpus to be too small for pre-training a language model. Consequently, we also used the tokenizer from the original Danish BERT model.

MeDa-BERT was pre-trained with a standard MLM objective where the model predicts the correct tokens of [MASK] tokens, as seen in Figure 3.6.

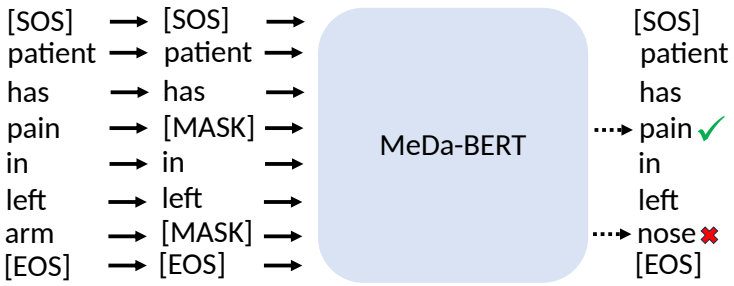


FIGURE 3.6: Illustration of masked language modelling.

By predicting the masked words, the model learns information about the context in which

the masked token appears. This allows the model to capture relationships and dependencies within a sentence. For pre-training MeDa-BERT, we corrupted 15% of tokens in the input sentence with the following procedure: 80% of the time we replaced a word with the [MASK] token, 10% of the time we replaced a word with a random word, and 10% of the time we did not replace the word. By doing this, the model is forced to consider and make good representations for all words in the input sentence as it does not know which words it will be asked to predict. Moreover, by keeping the word unchanged 10% of the time, the model is biased towards the actual word it should predict [4].

The model was pre-trained on a new Danish medical corpus that we collected from the internet and medical books. Acquiring this dataset was the main challenge for developing MeDa-BERT. For the English language, multiple databases of medical text are available, e.g., from PubMed [73] where text from millions of high-quality medical papers can be extracted. However, for the Danish language, no such database exists.

In the paper, we set out to investigate how much Danish medical text we could collect for pre-training language models. The final pre-training corpus contained approximately 134 million tokens, which is a significantly smaller corpus than English medical pre-trained language models [74, 75]. However, we showed that the corpus could be used to improve the performance of a Danish language model on four clinical classification tasks and surpass the performance of a general domain BERT model.

The four classification datasets were developed throughout my PhD-studies. The datasets were developed using EHRs from OUH and were annotated with a consensus label from three MDs. The datasets are described in detail in the paper.

The paper is presented below. It is licensed under a CC BY 4.0 license which can be found at `https://creativecommons.org/licenses/by/4.0/`.

Paper Correction: In Table 2 in the paper, the VTE label *Airways* should have been *Lower extremity*.

MeDa-BERT: A medical Danish pretrained transformer model

Jannik Skyttegaard Pedersen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
jasp@mmmi.sdu.dk

Martin Sundahl Laursen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
msla@mmmi.sdu.dk

Pernille Just Vinholt
Department of Clinical Biochemistry
Odense University Hospital

Thiusius Rajeeth Savarimuthu
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

Abstract

This paper introduces a medical Danish BERT-based language model (MeDa-BERT) and medical Danish word embeddings. The word embeddings and MeDa-BERT were pretrained on a new medical Danish corpus consisting of 133M tokens from medical Danish books and text from the internet. The models showed improved performance over general-domain models on medical Danish classification tasks. The medical word embeddings[1] and MeDa-BERT[2] are publicly available.

1 Introductions

Large language models (LLM) are powerful representation learners and have become the backbone structure of many modern natural language processing (NLP) systems. To learn text representations, LLM are first pretrained on a large-scale text corpus using self-supervised learning, e.g., masked language modelling. After pretraining, LLM are fine-tuned on specific downstream tasks where they have achieved state-of-the-art results on NLP benchmarks such as GLUE (Wang et al., 2018).

However, directly applying these general pretrained models to specialized domains such as the medical have led to unsatisfactory results (Peng et al., 2019). As a solution to this, a second round of in-domain pretraining (domain-adaptive pretraining) has shown to improve the performance of LLMs that were first trained on a general domain corpus (Gururangan et al., 2020). Domain-adaptive pretraining adjusts the weights of the LLM to better capture the terminology, style, and nuances that are relevant to the target domain.

Resource-rich languages such as English have large domain-specific corpuses available that have been used to develop e.g., biomedical (Lee et al., 2020), clinical (Alsentzer et al., 2019), scientific (Beltagy et al., 2019), and financial (Peng et al., 2021) LLMs that perform better than models trained on general corpuses. These models could potentially be used to improve human decision making, save time, and reduce costs, e.g., by extracting information from scientific articles, identifying potential drug interactions, and helping with NLP tasks such as text classification, named entity recognition, and question answering for each of their specialized domains.

For the Danish language, only LLMs trained on a general domain have been made publicly available[3]. This paper presents a medical Danish BERT model (MeDa-BERT)—a LLM trained on a new medical Danish text corpus. We also used the medical corpus to train medical word embeddings as they still have value in the clinical domain (Laursen et al., 2023). To evaluate the medical word embeddings and MeDa-BERT, we used existing medical Danish classification datasets. We found that an LSTM model using the medical word embeddings outperformed a similar model using general-domain word embeddings, and that MeDa-BERT performed slightly better than a general-domain BERT model.

2 Method

This section first describes how the medical corpus was collected and used to pretrain the medical Danish word embeddings and MeDa-BERT. Next, the datasets used to compare model performances and the fine-tuning procedure is described.

*Equal contribution

[1] https://huggingface.co/jannikskytt/MeDa-WE

[2] https://huggingface.co/jannikskytt/MeDa-Bert

[3] Pedersen et al. (2022b) developed a clinical transformer model but it is not publicly available

Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 301–307

Corpus	Type	Date retrieved	Tokens
Clinical guidelines	Guidelines	October - November 2022	80,567,576
Medicin.dk	Information portal	June 2021	28,878,335
FADL	Books	January 2022	12,531,373
Sundhed.dk	Information portal	May 2022	6,767,409
Netdoktor.dk	Information portal	October 2022	3,227,051
Wikipedia	Encyclopedia	October 2022	1,992,796
Total			**133,964,540**

Table 1: Number of tokens and date retrieved for each data source

2.1 Danish medical corpus

We collected data from the internet and from medical books. The owners of the data resources approved that we used their data in this study. We describe the data collection for each text contributor below. An overview of the text corpuses and their size can be seen in Table 1.

2.1.1 Clinical guidelines

We collected text from the document management systems of the five Danish regions. The documents contain guidelines and instructions for diagnostics and treatment of patients and all workflows that support this. The document systems also include non-medical documents from e.g. purchasing, logistics, and service departments which were removed. All departments that were excluded and the number of tokens retrieved from each region can be seen in Appendix A.

2.1.2 Medical information portals

We collected text from webpages that provide information to medical doctors and patients. The text was collected from Medicin.dk, Netdoktor.dk, and Sundhed.dk. The resources provide information about diseases, symptoms, and medical treatments. Moreover, the resources contain information specifically for health care professionals, e.g., medication guidelines and information about best practices in the field. Text not related to the medical domain and text written by non-professionals were removed from the corpus. A description of this process can be seen in appendix A.

2.1.3 Books

This part of the corpus consisted of 107 medical books from publisher FADLs Forlag that publishes books for medicine and nursing school.

2.1.4 Wikipedia

We used PetScan[4] to search for medical Wikipedia documents within predefined categories and its subcategories. We used a maximum depth of 5 for searching for subcategories. The following categories were used: anatomi, physiology, diseases, medication, epidemiology, diagnostics, medical procedures, medical specialities, medical physics, and medical equipment. We excluded documents with the categories: persons and companies. This process resulted in 5,391 documents. Next, we manually removed non-medical articles from that list which resulted in 5,266 documents.

2.2 Preprocessing of data

For all text corpusses, we defined a sample as one paragraph, i.e., a continuous stream of text without line breaks. We inserted spaces between alphanumeric and non-alphanumeric characters. Samples were further preprocessed to fit the pretraining procedure for either word embeddings or the transformer model, as detailed below.

2.2.1 Danish medical transformer model

MeDa-BERT was initialized with weights from a pretrained Danish BERT model[5] trained on 10.7 GB Danish text from Common Crawl (9.5 GB), Danish Wikipedia (221 MB), debate forums (168 MB), and Danish OpenSubtitles (881 MB).

For domain-adaptive pretraining, samples from the collected medical corpus were appended a [CLS] and [SEP] token in the start and end of each sample, respectively. Samples were concatenated to fit the maximum sequence length of 512 tokens and document boundaries were indicated by adding an extra [SEP] token in between samples. After this process, we removed duplicates corresponding to 0.2% of the total corpus. The model was trained using Adam (Kingma and Ba, 2015) with a weight decay of 0.01 as described in (Loshchilov and Hutter). Using gradient accumulation, the model was trained with a batch size of 4,032, a learning rate of 1e-4, and a linear learning rate decay warmed up over 1 epoch. The model was pretrained for a total of 48 epochs and evaluated after 16, 32, and 48 epochs. We used 5% of the samples as a validation set to evaluate the model during pretraining and trained the model on the remaining data using dynamic

[4] https://petscan.wmflabs.org/

[5] https://github.com/certainlyio/nordic_bert

Dataset	Label	Train	Validation	Test
Bleeding	Positive	10,331	1,300	1,300
	Negative	10,331	1,300	1,300
Bleeding site	Airways	1,000	125	125
	Cerebral	1,000	125	125
	Ear-nose-throat	1,000	125	125
	Eyes	1,000	125	125
	Gastrointestinal	1,000	125	125
	Gynecological	1,000	125	125
	Internal	1,000	125	125
	Skin	1,000	125	125
	Urogenital	1,000	125	125
	Unknown	1,000	125	125
VTE	Positive	9,064	1,100	1,100
	Negative	9,064	1,100	1,100
VTE site	Airways	1,600	200	200
	Lungs	1,600	200	200
	Unknown	1,600	200	200

Table 2: Dataset distributions

masked language modeling. The model was optimized using four Tesla v100 GPUs using the Huggingface (Wolf et al., 2020) library. All model parameters and pretraining losses are shown in Appendix B.

2.3 Danish medical word embeddings

We trained 300-dimensional FastText (Bojanowski et al., 2017) word embeddings. The embeddings were trained for 10 epochs using a window size of 5 and 10 negative samples. The hyperparameters were chosen to be able to compare the produced embeddings with the Danish FastText word embeddings from Grave et al. (2018) that were trained on a general domain.

2.4 Datasets

We compared performances between models using four medical datasets: bleeding classification, bleeding site classification, venous thromboembolism (VTE) classification, and VTE site classification. All samples were annotated with a consensus label from three medical doctors. The dataset distributions can be seen in Table 2 and examples of samples can be seen in Appendix C.

2.4.1 Bleeding classification

The bleeding dataset (Pedersen et al., 2021) is a binary classification problem with 25,862 samples. The dataset was constructed from 900 Danish electronic health records (EHR) from Odense University Hospital. The samples had an average token length of 13.3.

2.4.2 Bleeding site classification

The bleeding site dataset (Pedersen et al., 2022b) is a 10-class classification problem with 11,250 unique bleeding-positive samples annotated for the bleeding site. The bleeding site labels were: airways, cerebral, ear-nose-throat, eyes, gastrointestinal, gynecological, internal, skin, urogenital, and unknown. The dataset was constructed from 149,523 Danish EHR notes from Odense University Hospital. The samples had an average token length of 14.4.

2.4.3 VTE classification

The VTE dataset (Pedersen et al., 2022a) is a binary classification problem with 22,528 samples. The dataset was constructed from 94,520 Danish EHR notes from Odense University hospital. The samples had an average token length of 13.8.

2.4.4 VTE site classification

The VTE site dataset (Pedersen et al., 2022a) is a 3-class classification problem with 6,000 VTE-positive samples annotated for the VTE site. The VTE site labels were: airways, lungs, and unknown. The dataset was constructed from 94,520 Danish EHR notes from Odense University Hospital. The samples had an average token length of 14.5.

2.5 Fine-tuning

2.5.1 MeDa-BERT and BERT

We used the [CLS] token followed by a classification layer to classify samples of the datasets. We searched for the best models five times using Adam with learning rates [5e-5, 3e-5, 1e-5], i.e., we fine-tuned each model 15 times. The models were trained for a maximum of 10 epochs.

2.5.2 LSTM

We used the medical word embeddings as input to a bidirectional LSTM layer with a hidden layer size of 512. The last hidden state of the LSTM was followed by a dropout layer with probability 0.2, a dense layer of size 256, a ReLU activation function, a dropout layer of probability 0.2, and a dense classification layer. This model is referred to as LSTM+MeDa-WE.

The performance of the model is compared with another LSTM model (LSTM+General-WE) with the same parameters but using FastText embeddings trained on the general domain as input (Grave et al., 2018). We searched for the best

	Bleeding	Bleeding site	VTE	VTE site
LSTM+General-WE	$83.8_{.7}$	$69.3_{.8}$	$88.5_{.2}$	$86.4_{.7}$
LSTM+MeDa-WE	$\mathbf{91.4_{.3}}$	$\mathbf{84.9_{1.1}}$	$\mathbf{94.1_{.3}}$	$\mathbf{93.4_{.4}}$
BERT	$94.3_{.6}$	$86.7_{.8}$	$96.7_{.3}$	$94.7_{.3}$
MeDa-BERT_16	$94.7_{.3}$	$88.4_{.6}$	$\mathbf{97.1_{.4}}$	$95.5_{.2}$
MeDa-BERT_32	$95.1_{.5}$	$88.7_{.6}$	$96.9_{.3}$	$95.7_{.3}$
MeDa-BERT_48	$\mathbf{95.3_{.4}}$	$\mathbf{89.1_{.2}}$	$97.0_{.5}$	$\mathbf{95.8_{.3}}$

Table 3: Mean accuracy and standard deviation (subscript) for each model on four medical classification tasks. Best results for the LSTM and BERT-based models highlighted in bold. MeDa-BERT_16 denotes the MeDa-BERT model pretrained for 16 epochs.

models five times using Adam with learning rates [5e-5, 3e-5, 1e-5], i.e., we fine-tuned each model 15 times.

For all models we report the mean test set accuracy and standard deviation for the five best performing models on the validation dataset.

3 Results

Table 3 shows the results of each model on the four classification datasets.

3.1 Word embedding comparison

Using the medical word embeddings as input to an LSTM model resulted in large improvements compared to using general word embeddings. On average, LSTM+MeDa-WE outperformed the LSTM+General-WE model by 8.9 percentage points (PP). The largest improvement was seen on the 10-class bleeding site classification with an improvement of 15.6 PP.

3.2 Language model comparison

Comparing BERT and MeDa-BERT, the performance improvements were smaller. MeDa-BERT performed better on three of the datasets with an average improvement of 1.2 PP. The largest improvement was on the 10-class bleeding site classification with an improvement of 2.4 PP.

4 Discussion and limitations

This paper presented a new Danish medical corpus that was used to train NLP models. The corpus included medical books and text scraped from medical websites that provide information for both citizens and healthcare professionals. We applied different techniques to filter out non-medical data, e.g., by removing documents from non-medical departments or text written by non-healthcare professionals. While these steps did remove a large part of non-medical text, some non-medical text might still be present in the corpus. However, the results showed that models pretrained on the medical corpus performed better than general-domain models, especially for multiclass classification problems.

For the Danish language, few medical evaluation datasets are available and therefore the models were only evaluated on classification tasks. Moreover, the evaluation datasets were constructed from EHR text which has its own nuances compared to the text of the medical pretraining corpus, e.g., EHR text contains many spelling mistakes whereas the medical corpus contains few grammatical errors. These factors might limit the generalizability of the results. Future work should evaluate the models on other tasks, e.g., named-entity recognition and question answering which will provide a better understanding of the models' capabilities.

We found continuous small performance improvements by pretraining MeDa-BERT for more epochs. The model might improve with further pretraining but because of limited computational resources and the small rate of improvement, we did not explore this further. The model would also benefit from more medical pretraining data. Although this paper presented a large part of the available medical Danish text, more data could be collected, e.g., from other medical book publishers and websites.

The medical datasets used to evaluate the models are not publicly available because of privacy concerns. For future work, we will strive to publish parts of the medical corpus which requires permission from the text owners. We advise interested researchers to contact us for sharing possibilities.

5 Conclusion

This paper presented a Danish medical corpus consisting of 133M tokens. The corpus was used to pretrain medical word embeddings and language models. The models trained on the medical corpus performed better than similar models trained on a general domain.

Acknowledgement

We would like to thank the medical corpus contributors who gave acceptance of using their text in this project. Listed in alphabetical order: FADL's forlag, Medicin.dk, Netdoktor.dk, Sundhed.dk, The Capital Region of Denmark, The Central Region of Denmark, The Region of Northern Denmark, The Region of Southern Denmark, The Region of Zealand.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. https://doi.org/10.18653/v1/W19-1909 Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. https://doi.org/10.18653/v1/D19-1371 SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. https://aclanthology.org/Q17-1010 Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. https://doi.org/10.18653/v1/2020.acl-main.740 Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. http://arxiv.org/abs/1412.6980 Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Pernille Just Vinholt, Rasmus Søgaard Hansen, and Thiusius Rajeeth Savarimuthu. 2023. Benchmark for evaluation of danish clinical word embeddings. *Northern European Journal of Language Technology*, 9(1).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jannik Pedersen, Martin Laursen, Pernille Just, Anne Alnor, and Thiusius Savarimuthu. 2022a. Investigating anatomical bias in clinical machine learning algorithms. In *Findings of the European Chapter of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia. Association for Computational Linguistics.

Jannik S Pedersen, Martin S Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, et al. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and practice in thrombosis and haemostasis*, 5(4):e12505.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022b. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. https://doi.org/10.18653/v1/W19-5006 Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. https://doi.org/10.18653/v1/W18-5446 GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing*

and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6 Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendices

A Preprocessing of text corpuses

A.1 Medical information portals

Netdoktor.dk provides information about diseases, symptoms, medication, and treatment. Netdoktor.dk contains sections that are not related to the medical domain and discussion forums where users can communicate. Therefore, we removed documents having links containing the following strings: debat, kultur, testdigselv, behandlerguiden, nyhedsbrev, nyheder, privacypolicy, kontaktnetdoktor, cookieinformation, disclaimer, sponsorindhold and discussions. Moreover, citizens can ask medical questions [6] that are answered by health care professionals. We only included the answers to these questions.

Medicin.dk has three sub-pages: `www.min.medicin.dk` that provides information to citizens, `www.pro.medicin.dk` that provides information to health care professionals, and `www.indlaegssedler.dk` that contains information about medicine. We included all documents from these webpages.

Sundhed.dk provides information for medical professionals [7] and citizens [8] about diseases, symptoms, medication and treatment. We included all documents from these webpages.

A.2 Clinical guidelines

We collected clinical guidelines from the 5 regions of Denmark: The Capital Region of Denmark, The Region of Northern Denmark, The Region of Southern Denmark, The Region of Zealand, and The Central Region of Denmark. For each region we removed non-medical documents, seen in Table 4.

[6] `https://www.netdoktor.dk/brevkasser/`

[7] `https://www.sundhed.dk/sundhedsfaglig/`

[8] `https://www.sundhed.dk/borger/`

Region	Categories removed (in Danish)	Date retrieved	Tokens
Capital Region	Den sociale virksomhed Center for ejendomme Center for HR Center for Regional Udvikling Region Hovedstadens Apotek Steno Diabetes Center Copenhagena	October 2022	13,443,269
Northern Region	Logistik afdeling Teknisk Afdeling Himmerland Teknik Logistik Service	October 2022	6,505,559
Southern Region	Administration Service PsykInfo	September - November 2022	29,075,187
Region Zealand	Administration HR organisation og ledelse Indkøb IT PortørCentral Rengøring Økonomi Uddannelse	November 2022	6,387,083
Central Region		November 2022	25,156,478

Table 4: Categories removed and number of tokens from each of the Danish regions.

Parameter	Value
Architecture	
Number of layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Dropout	0.1
Attention dropout	0.1
Max seq. length	512
Optimization	
Learning rate	1e-4
Optimizer	AdamW
Adam weight decay	0.01
Adam epsilon	1e-6
Adam beta1	0.90
Adam beta2	0.98
Learning rate decay	Linear
Batch size	4032
Warm up	1 epoch
Epochs	16, 32, 48
Gradient clipping	1.0

Table 5: Architecture and optimization parameters for pretraining MeDa-BERT

B Model parameters and pretrainng loss

Table 5 shows the architecture and optimization parameters for pretraining MeDa-BERT. Table 6 shows the masked language modelling loss for MeDa-BERT during pretraining.

	Train loss	Validation loss
MeDa-BERT_16	2.122	2.019
MeDa-BERT_32	1.874	1.792
MeDa-BERT_48	1.766	1.673

Table 6: Masked language modelling loss for MeDa-BERT during pretraining. MeDa-BERT_16 denotes the model pretrained for 16 epochs.

C Dataset examples

Figure 7 shows a sample from each dataset translated from Danish to English.

Dataset	Example	Label
Bleeding	"Girl hospitalized on 14.05.11 with bleeding tendency. 1½ years ago, noticed bleeding on both arms, under the armpits and on the inner thighs. Subsequently blood discharges on the mucous membrane of the cheeks and quite heavy menstrual bleeding, which is unusual for pt."	Positive
Bleeding site	"19-year-old man referred by on-call doctor due to sudden onset of macroscopic hematuria and left-sided flank pain."	Urogenital
VTE	"Pt has severe heart failure and hence dyspnoea and the feeling of air hunger, and, in addition, pt has pulmonary embolisms and COPD. Treatment with Fragmin has started."	Positive
VTE site	"Irregular contours on the left side in the transverse sinus and beginning part of the sigmoid sinus compatible with partial thrombosis."	Brain

Table 7: Example of a sample from each dataset translated from Danish to English.

CHAPTER 4

Clinical bias mitigation

Dataset curation is the process of creating the datasets used to pre-train and fine-tune language models. Dataset curation plays an important role in developing safe language models as the data must be processed carefully to avoid or mitigate biases [10]. It includes acquiring EHRs; preprocessing; annotation; and splitting data into train, validation, and test sets. While bias can arise in any of these steps, this section describes a specific type of bias that arises due to the distribution of samples in the train, validation, and test splits used to train language models.

Our paper *Investigating anatomical bias in clinical classification algorithms* [17] stresses the importance of careful dataset curation when developing language models for clinical classification problems. During our research with bleeding extraction, we found large performance drops for specific patient populations if the samples of the dataset were not properly distributed between anatomical locations in the training dataset.

To validate our hypothesis, we created a controlled methodology to test the accuracy of an LSTM classifier and Clin-ELECTRA when the distribution of samples from different anatomical locations varied in the training dataset. We found that both classifiers were prone to anatomical bias in two clinical datasets and, therefore, we argued that researchers should carefully collect and construct their training, validation, and test data to include an equal distribution of samples from different anatomical locations for clinical classification problems.

Investigating anatomical bias in clinical machine learning algorithms

Jannik Skyttegaard Pedersen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
jasp@mmmi.sdu.dk

Martin Sundahl Laursen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
msla@mmmi.sdu.dk

Pernille Just Vinholt
Department of Clinical Biochemistry
Odense University Hospital

Anne Bryde Alnor
Department of Clinical Biochemistry
Odense University Hospital

Thiusius Rajeeth Savarimuthu
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

Abstract

Clinical machine learning algorithms have shown promising results and could potentially be implemented in clinical practice to provide diagnosis support and improve patient treatment. Barriers for realisation of the algorithms' full potential include bias which is systematic and unfair discrimination against certain individuals in favor of others.

The objective of this work is to measure *anatomical bias* in clinical text algorithms. We define anatomical bias as unfair algorithmic outcomes against patients with medical conditions in specific anatomical locations. We measure the degree of anatomical bias across two machine learning models and two Danish clinical text classification tasks, and find that clinical text algorithms are highly prone to anatomical bias. We argue that datasets for creating clinical text algorithms should be curated carefully to isolate the effect of anatomical location in order to avoid bias against patient subgroups.

1 Introduction

Research in clinical machine learning algorithms have shown promising results for automating clinical tasks. The algorithms could potentially be implemented in clinical practice to provide diagnosis support, improve patient treatment and provide time-savings for medical doctors (Topol, 2019; Matheny et al., 2020).

However, despite appealing research results, there are currently limited examples of algorithms being successfully deployed into clinical practice (Kelly et al., 2019). Barriers for realisation of the algorithms' full potential include bias and generalisation issues (Char et al., 2018; Hovy and Prabhumoye, 2021; Carrell et al., 2017).

Algorithmic bias can be defined as systematic and unfair discrimination against certain individuals or groups of individuals in favor of others (Friedman and Nissenbaum, 1996). Previous studies have raised serious concerns of algorithms that contain age, gender and racial bias (Sun et al., 2019; Davidson et al., 2019) — even for algorithms that have been taken into use (Obermeyer et al., 2019). Although machine learning algorithms are trained to be able to generalise to previously unseen data, they tend to overfit to the data they have been trained on. As a consequence of this, bias can unintendedly arise if some subgroups of the target population are not represented in the data used to train the algorithm. Moreover, if the training data itself include biases against some populations, e.g. data reflecting a negative attitude against people with disabilities (Hutchinson et al., 2020), these biases might be encoded and reinforced.

If biased algorithms are adopted, healthcare systems risk doing injustice to certain patient groups and harming patient safety (Obermeyer et al., 2019). Therefore, identifying and mitigating bias is important for successful implementation of novel clinical machine learning algorithms.

This paper investigates *anatomical bias* in clinical machine learning algorithms developed to classify and extract specific medical conditions from the narrative text of electronic health records (EHR). We define anatomical bias as unfair algorithmic outcomes against a subgroup of patients with the same medical condition, where the algorithm performs differently depending on the anatomical location of the condition. If the performance of clinical algorithms varies depending

*Equal contribution

Findings of the Association for Computational Linguistics: EACL 2023, pages 1398–1410

on the anatomical location, it is reflected in some patient subgroups receiving worse treatment than others.

We hypothesised that careful dataset curation is needed to measure and mitigate anatomical bias because the text description of medical conditions in EHRs varies depending on the location, e.g. 'epistaxis' is a location-specific word describing nose bleedings.

Specifically, this paper investigates anatomical bias for classification of bleeding and venous thromboembolism (VTE) mentions in the narrative text of Danish EHRs. Automatic extraction of these conditions could be valuable for medical doctors in clinical practice, e.g. to guide diagnostic decision making and treatment options (Decousus et al., 2011). Previous papers (Hinz et al., 2013; Lee et al., 2017; Taggart et al., 2018; Li et al., 2019; Mitra et al., 2020, 2021; Elkin et al., 2021; Pedersen et al., 2021; Shi et al., 2021; Verma et al., 2022) have shown promising results for automatic extraction of these medical conditions but they did not investigate the performance of the algorithms across anatomical subgroups.

Our main contributions are:

- We find that clinical text algorithms are highly prone to anatomical bias.
- The performance of state-of-the-art algorithms developed to extract specific medical conditions varies significantly across anatomical locations with performance drops up to 89.1 percentage points (PP).
- We argue that datasets for creating clinical text algorithms should be curated carefully to isolate the effect of anatomical location in order to avoid bias against patient subgroups.

2 Methods

To investigate if machine learning algorithms are prone to anatomical bias, we performed two experiments. We (1) investigated the performance of a binary classifier on different anatomical subgroups of a medical condition when that subgroup was left out of the training set, and (2) measured how the performance on an anatomical subgroup varied depending on the amount of samples from that subgroup included in the training set.

Table 1: Distribution of the training, validation, and test samples for the balanced bleeding detection dataset.

Label	Location	Train	Validation	Test
Positive for bleeding	Gastrointestinal	750	250	250
	Urogenital	750	250	250
	Internal	750	250	250
	Otorhinolaryngeal	750	250	250
	Dermatological	750	250	250
	Gynecological	750	250	250
	Cerebral	750	250	250
	Ophthalmological	750	250	250
Negative for bleeding		6,000	2,000	2,000
Sum		12,000	4,000	4,000

Table 2: Distribution of the training, validation, and test samples for the balanced VTE detection dataset.

Label	Location	Train	Validation	Test
Positive for VTE	Lower extremity	1,600	200	200
	Lung	1,600	200	200
	Liver	0	0	239
	Cerebral	0	0	218
	Upper extremity	0	0	176
Negative for VTE		3,200	400	1,033
Sum		6,400	800	2,066

2.1 Datasets

We used the binary bleeding classification dataset from Pedersen et al. (2022) and present a new binary VTE classification dataset. The bleeding dataset consists of 20,000 sentences from Danish EHRs labeled as either positive or negative for bleeding mentions. The VTE classification dataset consists of 9,266 sentences from Danish EHRs labeled as either positive or negative for VTE mentions. Both datasets were constructed from Danish EHRs from Odense University Hospital and were labeled with a consensus label from three medical doctors.

In addition to the main labels of each dataset (positive and negative for bleeding or VTE), we created a subgroup label for the positive samples describing the anatomical location of either the bleeding or VTE mention. Samples that did not describe the anatomical location or described multiple locations were omitted.

For the bleeding dataset, we used the following eight anatomical locations: gastrointestinal, urogenital, internal, otorhinolaryngeal, dermatological, gynecological, cerebral, and ophthalmological.

For the VTE dataset, we used the following five anatomical locations: lower extremity, lung, liver, cerebral, and upper extremity.

The locations included for each medical condition were selected by two medical doctors.

For each dataset, we created a balanced training, validation, and test set containing an equal amount

of positive and negative samples. Moreover, for the bleeding dataset, the positive samples of the training, validation, and test sets were distributed equally between anatomical locations. For the VTE dataset, only samples from the lower extremity and lung locations were distributed equally between the train, validation, and test sets. The liver, cerebral, and upper extremity locations were only used for the test set because of a limited number of samples.

All samples were preprocessed by removing special characters, superfluous spaces, and duplicate samples. After preprocessing the samples, the bleeding and VTE datasets had an average token length of 13.3 and 13.6, respectively. The dataset distributions can be seen in Table 1 and Table 2.

2.2 Training set distributions

To measure performance differences for specific anatomical locations, we systematically removed all samples from a specific location, x, from the training set, creating the training set $\mathcal{T}_{\not\in x}$, trained a deep learning model on $\mathcal{T}_{\not\in x}$, and evaluated it on the test set. For example, for the bleeding dataset, we created 8 different training sets, one for each anatomical location being removed, containing 10,500 samples.

For comparison, we created a balanced training set, $\mathcal{T}$, which included the same amount of samples as $\mathcal{T}_{\not\in x}$, distributed equally between the positive and negative classes, and between anatomical locations.

2.3 Deep learning models

The deep learning models were a transformer-based ELECTRA model (Clark et al., 2020) and a Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997).

The ELECTRA model was a Danish clinical ELECTRA (Clin-ELECTRA) (Pedersen et al., 2022) pretrained on the narrative text from 299,718 EHRs from Odense University Hospital. The model had ∼13M parameters and consisted of 12 transformer layers with 4 attention heads. We initialised Clin-ELECTRA from its pretrained checkpoint and followed the HuggingFace (Wolf et al., 2019) implementations for binary text classification.

The LSTM model had ∼4M parameters and consisted of a bidirectional LSTM layer with a hidden layer size of 512. The last hidden state of the LSTM was followed by a dropout layer with probability 0.2, a dense layer of size 256, a ReLU activation function, a dropout layer of probability 0.2, and a dense classification layer. For word representation, the LSTM model used 300-dimensional FastText (Bojanowski et al., 2017) word embeddings pretrained on Danish EHRs consisting of 1.4B tokens.

2.4 Model evaluation

For each of the ELECTRA and LSTM deep learning models and training sets $\mathcal{T}$ and $\mathcal{T}_{\not\in x}$, we:

1. Trained the deep learning model with five different learning rates and random initialisations.
2. Computed the test set accuracy of the best performing model based on the loss on the validation set.
3. Repeated step 1 and 2 five times.

We used the five accuracies to perform bootstrapping with 9,999 replicates and calculated mean accuracy, standard error (SE), and 95% confidence interval (CI) for $\mathcal{T}$ and $\mathcal{T}_{\not\in x}$. Moreover, we computed the bootstrapped difference of means between $\mathcal{T}$ and $\mathcal{T}_{\not\in x}$ to evaluate statistically significant differences in performance.

For both deep learning models, we used the Adam optimizer (Kingma and Ba, 2015) and searched for the best model using learning rates 7e-5, 8e-5, 9e-5, and 1e-4. Clin-ELECTRA was fine-tuned for a maximum of 10 epochs and the LSTM for a maximum of 30 epochs. One epoch was trained in <1 and ∼5 seconds for the LSTM and ELECTRA model, respectively, using an NVIDIA v100 GPU.

We measured anatomical bias as the difference in sensitivity on a specific location, x, between two deep learning models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\in x}$.

3 Results

3.1 Bleeding classification

Table 3 shows the binary accuracy of the bleeding classifiers on the test set for each of the training sets $\mathcal{T}$ and $\mathcal{T}_{\not\in x}$. Appendix A shows additional metrics. With the exception of $\mathcal{T}_{\not\in Otorhinolaryngeal}$ for the ELECTRA model, all training sets with an anatomical location removed resulted in models with significantly worse performance than when trained on $\mathcal{T}$.

The decreases in accuracy were caused by a significant drop in sensitivity for the anatomical locations which had been removed from the training

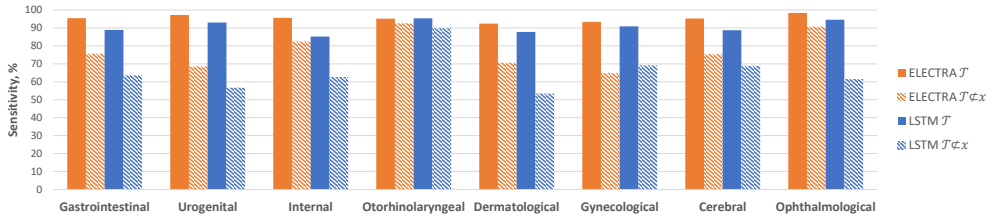


Figure 1: Sensitivity of models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ for each anatomical location, x.

Table 3: Accuracy (%), standard error (SE), and 95% confidence interval (CI) for the bleeding classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$.

	ELECTRA		LSTM	
	Accuracy±SE	**CI**	**Accuracy±SE**	**CI**
$\mathcal{T}$	95.6 ± 0.1	95.4 - 95.8	90.6 ± 0.1	90.4 - 90.7
$\mathcal{T}_{\not\subset Gastrointestinal}$	94.4 ± 0.2*	94.0 - 94.7	88.8 ± 0.1*	88.6 - 88.9
$\mathcal{T}_{\not\subset Urogenital}$	93.9 ± 0.1*	93.7 - 94.2	88.5 ± 0.1*	88.3 - 88.6
$\mathcal{T}_{\not\subset Internal}$	95.0 ± 0.1*	94.8 - 95.2	88.8 ± 0.1*	88.5 - 89.0
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	95.6 ± 0.1	95.4 - 95.8	90.0 ± 0.2*	89.8 - 90.4
$\mathcal{T}_{\not\subset Dermatological}$	94.3 ± 0.1*	94.0 - 94.5	88.2 ± 0.1*	88.1 - 88.3
$\mathcal{T}_{\not\subset Gynecological}$	93.8 ± 0.1*	93.6 - 94.0	89.1 ± 0.1*	88.9 - 89.4
$\mathcal{T}_{\not\subset Cerebral}$	94.6 ± 0.2*	94.3 - 94.9	89.0 ± 0.1 *	88.8 - 89.2
$\mathcal{T}_{\not\subset Ophthalmological}$	95.3 ± 0.1*	95.1 - 95.4	88.2 ± 0.2 *	87.8 - 88.5

Table 4: Accuracy (%), standard error (SE), and 95% confidence interval (CI) for the VTE classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$.

	ELECTRA		LSTM	
	Accuracy±SE	**CI**	**Accuracy±SE**	**CI**
$\mathcal{T}$	84.8 ± 0.3	84.2 - 85.4	75.9 ± 0.3	75.4 - 76.4
$\mathcal{T}_{\not\subset Lower\ extremity}$	67.6 ± 0.6*	66.5 - 68.7	71.6 ± 0.4*	70.8 - 72.6
$\mathcal{T}_{\not\subset Lung}$	74.0 ± 1.1*	71.9 - 76.1	69.9 ± 0.1*	69.7 - 70.2

data. Figure 1 shows the test set sensitivity for each anatomical location, x, for each training set $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. The sensitivity for all anatomical locations was significantly worse when not present in the training set with performance drops up to 28.8 PP for ELECTRA and 36.3 PP for the LSTM model. On average, the sensitivity on the anatomies decreased with 17.8 PP (standard deviation ± 8.8 PP) for ELECTRA and 24.5 PP (standard deviation ± 9.4 PP) for the LSTM model.

Moreover, it is seen that even though models trained on $\mathcal{T}_{\not\subset x}$ achieved high accuracies on the test set overall, the sensitivity on the anatomical location not present in the training set was low. E.g., for ELECTRA, $\mathcal{T}_{\not\subset Gynecological}$ had a 93.8% accuracy on the test set, but the sensitivity for gynecological bleedings was only 64.8%. Appendix A shows the sensitivity, SE, and the differences of means for all anatomical locations and training sets.

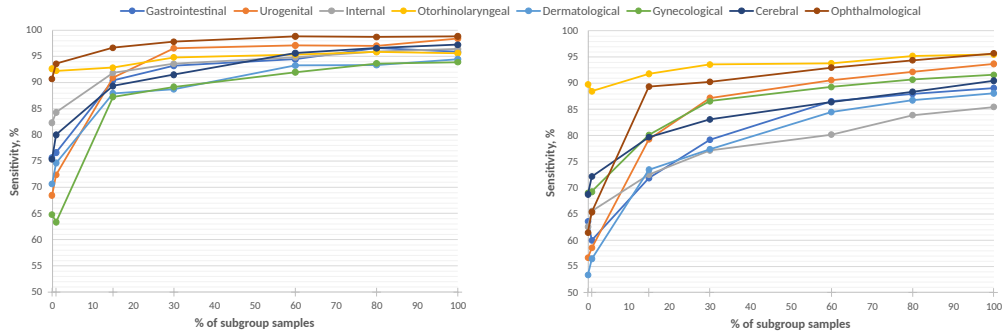
Figure 2 shows the sensitivity on each anatomical location by the percentage of total subgroup samples in the training set. It is seen that the accuracy increased as more samples were present in the training set. For the LSTM model, the sensitivity on gastrointestinal, urogenital, cerebral, and ophthalmological bleedings was significantly worse - even when 80% of samples were present in the training set. For ELECTRA, the sensitivity on urogenital and internal bleedings was significantly worse when 80% of samples were present in the training set. Appendix A shows the accuracies and differences of means.

3.2 Venous thromboembolism classification

Table 4 shows the binary accuracy of the VTE classifiers on the test set for each of the training sets $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. Appendix B shows additional metrics. Models trained on $\mathcal{T}_{\not\subset Lower\ extremity}$ and $\mathcal{T}_{\not\subset Lung}$ performed significantly worse than those trained on $\mathcal{T}$.

Similar to the bleeding classifier results, the decrease in the overall accuracy was caused by a significant drop in sensitivity on the anatomical locations which had been removed from the training data. Figure 3 shows the sensitivity for each anatomical location, x, for each training set $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. The sensitivity on liver, cerebral, and lower extremity VTEs is only reported when not being present in the training set because of limited samples.

The sensitivity on lower extremity and lung VTEs was significantly worse when not present in the training set, e.g. the performance for the ELECTRA classifier decreased by 89.1 PP for lower extremity VTEs and 81.0 PP for lung VTEs. Appendix B shows the sensitivity, SE, and the differences of means for all anatomical locations and



(a) ELECTRA (b) LSTM

Figure 2: Test set sensitivity on the anatomical locations when removing a fraction of samples from that anatomy from the training set.

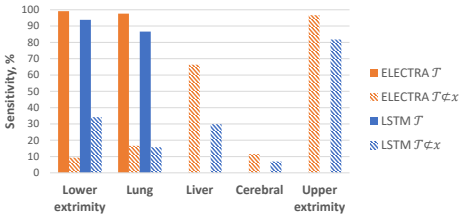


Figure 3: Sensitivity of models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ for each anatomical location, x. The sensitivity on liver, cerebral, and lower extremity VTEs is only reported when not being present in the training set because of limited samples.

training sets.

Figure 4 shows the sensitivity for lower extremity and lung VTEs by the percentage of total subgroup samples in the training set. Both locations performed significantly worse when 15% and 30% of samples from that location were present in the training set for the ELECTRA and LSTM classifier, respectively. Appendix B shows the accuracies and differences of means.

4 Analysis of word distributions

Medical conditions are often described using different words depending on the anatomical location for which the condition occurs. Table 5 shows the top-3 most frequent words used to describe VTE events for each anatomical location. The column *Location uniqueness* shows the fraction of times a word appears in samples from a specific anatomical location compared to the complete dataset:

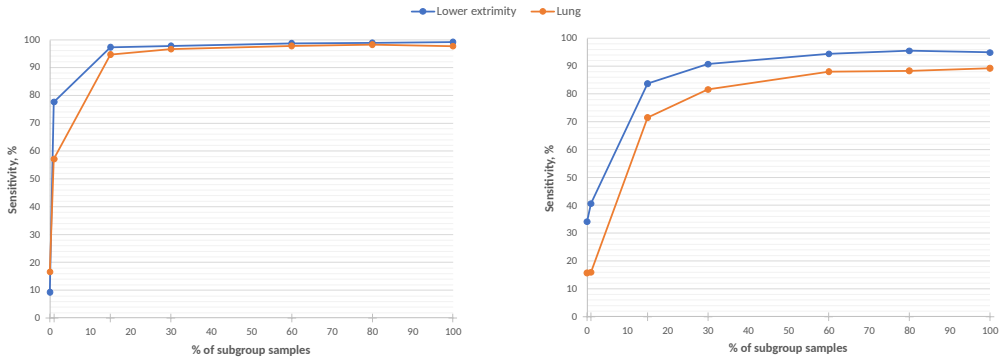
$$Location\ uniqueness = \frac{f_x}{f_D} \quad (1)$$

where f_x is the frequency of the word in samples from anatomical location x and f_D is the total frequency of the word in the dataset, i.e. a value of 1 means that the word is unique for an anatomical location.

The top-3 words for upper extremity had a low uniqueness score (<0.15). This indicates that the vocabulary used to describe VTEs in the upper extremity was also used for other locations — e.g.

Table 5: Most frequent words used to describe VTE events for each anatomical location of the VTE classification dataset. Words are translated from Danish to English and, therefore, some cells include two words. PE = pulmonary embolism.

Word	**Frequency**	**Location uniqueness**
Location: Lower extremity		
dvt	1384	0.92
thrombus	135	0.70
blood clot	108	0.47
Location: Lung		
pulmonary embolism	1058	0.98
le (PE)	483	0.99
pulmonary embolisms	242	0.99
Location: Liver		
porta thrombosis	71	1.00
thrombosis	70	0.36
thrombus	22	0.11
Location: Cerebral		
infarct	93	0.97
sinus thrombosis	32	0.97
blood clot	26	0.11
Location: Upper extremity		
dvt	99	0.07
thrombus	28	0.14
thrombosis	14	0.07



(a) ELECTRA (b) LSTM

Figure 4: Test set sensitivity on the anatomical locations when removing a fraction of samples from that anatomy from the training set.

'dvt' (deep vein thrombosis) was the most frequent word but the uniqueness score was only 0.07. This might explain why the sensitivity for upper extremity was relatively high, as seen in Figure 3, even when samples from that location were not present in the training set. On the contrary, some of the frequent words from the lower extremity, lung, and cerebral locations were close to unique which could explain why the sensitivity of those locations were low. Appendix C shows word frequency and location uniqueness for the bleeding classification dataset which shows similar results.

5 Discussion

This paper has presented evidence that clinical natural language processing (NLP) classification algorithms are prone to anatomical bias, which is unfair algorithmic outcomes against patients with medical conditions in specific anatomical locations. We found that the performance of algorithms for both bleeding and VTE classification can vary significantly depending on the anatomical location with differences up to 36.3 PP and 89.1 PP, respectively.

Moreover, we found that small fluctuations in the training set distribution of anatomical locations can lead to significant performance drops for the underrepresented anatomical locations. For the datasets presented in this study, we showed that the words used to describe medical conditions vary depending on the anatomical location. If classifiers do not learn to properly represent the full vocabulary for describing a medical condition, its performance will decrease for some anatomical locations.

We argue that datasets for clinical NLP algorithms should be created to be able to carefully measure anatomical bias, e.g. by subdividing each sample into an anatomical location. This is essential to avoid implementing clinical algorithms that might discriminate against specific subgroups of patients. For example, one of the developed VTE classifiers in this study performed with sensitivities of >96% for VTEs in the lungs and lower extremity while it performed with a sensitivity of only 11.5% for cerebral VTEs. Applying such a model in clinical practice or research would provide unfair algorithmic outcomes against patients with cerebral VTEs. We also showed that an algorithm not exposed to gynecological bleedings would perform worse on this anatomical location. This would lead to unfair algorithmic outcomes against woman with gynecological bleedings. Similarly, because alcoholics have an increased prevalence of gastrointestinal bleedings (Singal et al., 2018), this group of people would have a higher risk of unfair algorithmic outcomes if the algorithm has not been trained on such bleeding locations.

To the best of our knowledge, anatomical bias has not been investigated in previous research. However, some studies tried to automatically create datasets distributed between different patient groups by extracting data based on International Classification of Diseases 10 (ICD) codes — e.g. Pedersen et al. (2021) extract data based on different bleeding disorders. While this approach could, to some degree, mitigate the problem, studies (Valkhoff et al., 2014; Øie et al., 2018) found that ICD codes have low accuracy and, therefore, this does not ensure an evenly distributed dataset.

Moreover, in order to isolate and measure the performance on different anatomical locations, the dataset should be constructed with a known distribution of these anatomies.

Our work is closely related to the field of domain adaption. For example, MacAvaney et al. (2017) find that an algorithm trained to extract temporal information from a specific patient population performs worse on another related patient population. Their results highlight that it is a challenging task to develop algorithms that can generalise well across domains. The main difference between our study and theirs is that the algorithms described in this paper are not developed to work on different domains. Rather, the algorithms are specifically developed to work on a specialised domain in the clinical field, e.g. bleeding detection. As our results have shown, the algorithms perform worse on some subpopulations of the population it is supposed to work on, and therefore, we describe this as a bias issue.

6 Conclusion

This paper presented evidence that clinical NLP algorithms are prone to anatomical bias. We found that the performance of clinical classification algorithms for both bleeding and VTE classification can vary significantly depending on the anatomical location of the medical condition. We argue that anatomical bias should be carefully examined when developing clinical text algorithms in order to avoid unfair algorithm performance against patient subgroups.

7 Limitations

Future work should investigate the degree of anatomical bias in other clinical areas and tasks, e.g. named entity recognition, to be able to compare the severity of the bias problem between algorithms and other clinical areas. Moreover, as the datasets used in this study are only from a single institution, the findings of the paper might not be widely representative.

The objective of this work was to stress the need for measuring anatomical bias. We leave it to future work to investigate algorithmic solutions other than dataset balancing for mitigating the problem, e.g. using techniques such as oversampling and data augmentation. Such techniques could also help mitigating anatomical bias in algorithms for which training set balancing is not sufficient.

The classification datasets and machine learning models presented in this paper cannot be shared publicly due to privacy concerns but we advise interested researchers to contact us for sharing possibilities.

Ethics Statement

Machine learning researchers must be proactive in recognising and counteracting biases such as the one described in this paper. We hope that the findings and focus of this paper will lead other researchers to test and mitigate other kinds of algorithmic biases.

All datasets used in this research were obtained according to each dataset's respective data usage policy. The datasets were stored and processed on a secure platform[1] in compliance with GDPR regulations. According to section 14(2) of the Danish Act on Research Ethics Review of Health Research Projects[2], studies using retrospective data that do not involve human biological material do not require ethical approval.

References

Piotr Bojanowski, Édouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

David S Carrell, Robert E Schoen, Daniel A Leffler, Michele Morris, Sherri Rose, Andrew Baer, Seth D Crockett, Rebecca A Gourevitch, Katie M Dean, and Ateev Mehrotra. 2017. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991.

Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

[1] https://docs.cloud.sdu.dk/intro/security.html

[2] https://www.retsinformation.dk/eli/lta/2011/593 (English version is unfortunately not available)

Hervé Decousus, Victor F Tapson, Jean-François Bergmann, Beng H Chong, James B Froehlich, Ajay K Kakkar, Geno J Merli, Manuel Monreal, Mashio Nakamura, Ricardo Pavanello, et al. 2011. Factors at admission associated with bleeding risk in medical patients: findings from the improve investigators. *Chest*, 139(1):69–79.

Peter L Elkin, Sarah Mullin, Jack Mardekian, Christopher Crowner, Sylvester Sakilay, Shyamashree Sinha, Gary Brady, Marcia Wright, Kimberly Nolen, JoAnn Trainer, et al. 2021. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *Journal of medical Internet research*, 23(11):e28946.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347.

Eugenia R McPeek Hinz, Lisa Bastarache, and Joshua C Denny. 2013. A natural language processing algorithm to define a venous thromboembolism phenotype. In *AMIA Annual Symposium Proceedings*, volume 2013, page 975. American Medical Informatics Association.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hee-Jin Lee, Min Jiang, Yonghui Wu, Christian M Shaffer, John H Cleator, Eitan A Friedman, Joshua P Lewis, Dan M Roden, Josh Denny, and Hua Xu. 2017. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. *AMIA Summits on Translational Science Proceedings*, 2017:185.

Rumeng Li, Baotian Hu, Feifan Liu, Weisong Liu, Francesca Cunningham, David D McManus, Hong Yu, et al. 2019. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. *JMIR medical informatics*, 7(1):e10788.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. Guir at semeval-2017 task 12: a framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029.

Michael E Matheny, Danielle Whicher, and Sonoo Thadaney Israni. 2020. Artificial intelligence in health care: a report from the national academy of medicine. *Jama*, 323(6):509–510.

Avijit Mitra, Bhanu Pratap Singh Rawat, David McManus, Alok Kapoor, and Hong Yu. 2020. Bleeding entity recognition in electronic health records: A comprehensive analysis of end-to-end systems. In *AMIA Annual Symposium Proceedings*, volume 2020, page 860. American Medical Informatics Association.

Avijit Mitra, Bhanu Pratap Singh Rawat, David D McManus, Hong Yu, et al. 2021. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. *JMIR medical informatics*, 9(7):e27527.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Lise R Øie, Mattis A Madsbu, Charalampis Giannadakis, Anders Vorhaug, Heidi Jensberg, Øyvind Salvesen, and Sasha Gulati. 2018. Validation of intracranial hemorrhage in the norwegian patient registry. *Brain and behavior*, 8(2):e00900.

Jannik S Pedersen, Martin S Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, et al. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and practice in thrombosis and haemostasis*, 5(4):e12505.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Jianlin Shi, John F Hurdle, Stacy A Johnson, Jeffrey P Ferraro, David E Skarda, Samuel RG Finlayson, Matthew H Samore, and Brian T Bucher. 2021. Natural language processing for the surveillance of

postoperative venous thromboembolism. *Surgery*, 170(4):1175–1182.

Ashwani K Singal, Ramon Bataller, Joseph Ahn, Patrick S Kamath, and Vijay H Shah. 2018. Acg clinical guideline: alcoholic liver disease. *The American journal of gastroenterology*, 113(2):175.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Maxwell Taggart, Wendy W Chapman, Benjamin A Steinberg, Shane Ruckel, Arianna Pregenzer-Wenzler, Yishuai Du, Jeffrey Ferraro, Brian T Bucher, Donald M Lloyd-Jones, Matthew T Rondina, et al. 2018. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA network open*, 1(6):e183451–e183451.

Eric Topol. 2019. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

Vera E Valkhoff, Preciosa M Coloma, Gwen MC Masclee, Rosa Gini, Francesco Innocenti, Francesco Lapi, Mariam Molokhia, Mees Mosseveld, Malene Schou Nielsson, Martijn Schuemie, et al. 2014. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *Journal of clinical epidemiology*, 67(8):921–931.

Amol A Verma, Hassan Masoom, Chloe Pou-Prom, Saeha Shin, Michael Guerzhoy, Michael Fralick, Muhammad Mamdani, and Fahad Razak. 2022. Developing and validating natural language processing algorithms for radiology reports compared to icd-10 codes for identifying venous thromboembolism in hospitalized medical patients. *Thrombosis Research*, 209:51–58.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Bleeding classification results

Table 6: Precision, recall, and F1 performance for the bleeding classification dataset.

Table 7: Sensitivity and standard error for all anatomical locations of the bleeding classification dataset.

Table 8: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the bleeding classification dataset.

Table 9: Bleeding test set sensitivity and standard error on an anatomical location by percentage of subgroup samples in the modified training set.

Table 10: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the bleeding classification dataset.

B VTE classification results

Table 11: Precision, recall, and F1 performance for the VTE classification dataset.

Table 12: Sensitivity and standard error for all anatomical locations of the VTE classification dataset.

Table 13: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the VTE classification dataset.

Table 14: VTE test set sensitivity and standard error on an anatomical location by percentage of subgroup samples in the modified training set.

Table 15: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the VTE classification dataset.

C Bleeding word distribution

Table 16: Most frequent words used to describe bleeding mentions for each anatomical location of the bleeding classification dataset.

Table 6: Precision, recall, and F1 performance for the bleeding classification dataset. $\mathcal{T}_{\not\in x}$ denotes the training set from which an anatomical location, x, has been removed. SE = Standard error. CI = 95% confidence interval.

	ELECTRA			LSTM		
	Precision ± SE (CI)	**Recall ± SE (CI)**	**F1 ± SE (CI)**	**Precision ± SE (CI)**	**Recall ± SE (CI)**	**F1 ± SE (CI)**
$\mathcal{T}_{\not\in Gastrointestinal}$	94.7 ± 0.2 (94.2 - 95.0)	94.2 ± 0.4 (93.4 - 95.0)	94.4 ± 0.2 (94.0 - 94.8)	89.1 ± 0.7 (8.77 - 90.5)	88.5 ± 0.7 (87.2 - 89.8)	88.8 ± 0.1 (88.6 - 88.9)
$\mathcal{T}_{\not\in Urogenital}$	94.3 ± 0.3 (93.6 - 94.8)	93.5 ± 0.4 (92.6 - 94.3)	93.9 ± 0.1 (93.6 - 94.2)	89.5 ± 0.1 (89.3 - 89.8)	87.3 ± 0.2 (87.0 - 87.7)	88.5 ± 0.1 (88.3 - 88.6)
$\mathcal{T}_{\not\in Internal}$	95.0 ± 0.4 (94.2 - 95.0)	95.1 ± 0.6 (93.8 - 96.1)	95.0 ± 0.1 (94.8 - 95.2)	89.9 ± 0.4 (89.3 - 90.7)	87.6 ± 0.5 (86.6 - 88.6)	88.8 ± 0.1 (88.5 - 89.0)
$\mathcal{T}_{\not\in Otorhinolaryngeal}$	95.2 ± 0.2 (94.8 - 95.6)	96.2 ± 0.1 (96.0 - 96.3)	95.7 ± 0.1 (95.5 - 95.8)	89.7 ± 0.3 (89.1 - 90.3)	90.7 ± 0.6 (89.6 - 91.7)	90.1 ± 0.2 (89.8 - 90.5)
$\mathcal{T}_{\not\in Dermatological}$	94.7 ± 0.5 (93.6 - 95.5)	93.7 ± 0.4 (93.1 - 94.5)	94.2 ± 0.1 (94.0 - 94.4)	89.5 ± 0.3 (89.0 - 90.2)	87.0 ± 0.3 (86.5 - 87.6)	88.2 ± 0.1 (88.1 - 88.3)
$\mathcal{T}_{\not\in Gynecological}$	94.2 ± 0.2 (93.7 - 94.6)	93.5 ± 0.2 (93.1 - 93.8)	93.8 ± 0.1 (93.6 - 94.0)	89.6 ± 0.1 (89.4 - 89.8)	88.7 ± 0.2 (88.3 - 89.0)	89.1 ± 0.1 (88.9 - 89.4)
$\mathcal{T}_{\not\in Cerebral}$	95.2 ± 0.2 (94.7 - 95.6)	94.0 ± 0.2 (93.5 - 94.5)	94.6 ± 0.1 (94.3 - 94.8)	89.1 ± 0.2 (88.7 - 89.5)	88.9 ± 0.2 (88.6 - 89.4)	89.0 ± 0.1 (88.8 - 89.2)
$\mathcal{T}_{\not\in Ophthalmological}$	95.0 ± 0.1 (94.7 - 95.3)	95.6 ± 0.2 (95.2 - 96.0)	95.3 ± 0.1 (95.1 - 95.5)	88.8 ± 0.4 (88.0 - 89.6)	87.7 ± 0.5 (87.0 - 88.8)	88.2 ± 0.2 (87.8 - 88.5)
$\mathcal{T}$	95.9 ± 0.2 (95.5 - 96.2)	95.4 ± 0.3 (94.7 - 96.0)	95.6 ± 0.1 (95.4 - 95.8)	90.5 ± 0.5 (89.6 - 91.4)	90.4 ± 0.6 (89.3 - 91.7)	90.5 ± 0.1 (90.3 - 90.6)

Table 7: Sensitivity (%) and standard error for all anatomical locations of the bleeding classification dataset. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}_{\not\in x}$ and $\mathcal{T}$.

	Gastrointestinal	**Urogenital**	**Internal**	**Otorhinolaryngeal**	**Dermatological**	**Gynecological**	**Cerebral**	**Ophthalmological**
	ELECTRA							
$\mathcal{T}_{\not\in Gastrointestinal}$	75.7± 2.4*	98.6± 0.3*	96.2± 0.3	96.2± 0.4*	94.3± 0.4*	95.9± 0.4*	97.3± 0.3*	99.1± 0.1*
$\mathcal{T}_{\not\in Urogenital}$	97.1±0.3*	68.5± 2.2*	97.4± 0.1*	96.6± 0.2*	95.8± 0.6*	95.8± 0.5*	98.1± 0.1*	98.7± 0.3
$\mathcal{T}_{\not\in Internal}$	96.6±0.2*	98.8± 0.2*	82.3± 2.0*	96.2± 0.3*	94.6± 0.6*	95.6± 1.0*	97.3± 0.3*	99.0± 0.2*
$\mathcal{T}_{\not\in Otorhinolaryngeal}$	96.2±0.4	98.2± 0.2*	96.2± 0.3	92.6± 0.4*	94.6± 0.2*	95.1± 0.5*	97.4± 0.1*	99.1± 0.2*
$\mathcal{T}_{\not\in Dermatological}$	96.6±0.4	98.2± 0.3*	96.2± 0.4*	95.8± 0.5	70.6± 1.4*	96.0± 0.7*	97.6± 0.2*	98.7± 0.2
$\mathcal{T}_{\not\in Gynecological}$	97.0±0.2*	98.6± 0.3*	97.2± 0.3*	97.2± 0.2*	95.5± 0.4*	64.8± 0.8*	97.9± 0.4*	99.4± 0.1*
$\mathcal{T}_{\not\in Cerebral}$	96.2±0.3	98.3± 0.3*	96.6± 0.4	99.1± 0.1*	94.4± 0.5*	95.8± 0.4*	75.4± 0.7*	99.1± 0.1*
$\mathcal{T}_{\not\in Ophthalmological}$	96.6±0.2*	98.5± 0.2*	96.5± 0.2*	95.2± 0.3	93.8± 0.6	95.4± 0.6*	97.4± 0.3*	90.7± 0.5*
$\mathcal{T}$	95.4± 0.4	97.3± 0.4	95.6± 0.5	95.2± 0.2	92.4± 0.8	93.4± 0.2	95.3± 0.4	98.4± 0.3
	LSTM							
$\mathcal{T}_{\not\in Gastrointestinal}$	63.6 ± 3.0*	94.6 ± 0.7*	85.8 ± 1.0*	96.1 ± 0.5*	88.2 ± 0.9	92.2 ± 0.8	90.6 ± 0.7*	96.6 ± 0.6*
$\mathcal{T}_{\not\in Urogenital}$	90.2 ± 0.4	56.7 ± 1.6*	86.0 ± 0.5	97.4 ± 0.2*	89.8 ± 0.3*	91.6 ± 0.4	90.9 ± 0.3*	96.2 ± 0.1
$\mathcal{T}_{\not\in Internal}$	86.9 ± 0.6	93.0 ± 0.7	62.6 ± 1.2*	95.1 ± 0.5	86.9 ± 0.5	89.7 ± 0.4	90.6 ± 0.5	96.1 ± 0.4
$\mathcal{T}_{\not\in Otorhinolaryngeal}$	88.5 ± 0.6	93.8 ± 0.6	85.2 ± 1.0	89.8 ± 0.5*	89.4 ± 0.8	92.4 ± 0.7	90.2 ± 0.5*	96.1 ± 0.4*
$\mathcal{T}_{\not\in Dermatological}$	90.6 ± 0.3	93.4 ± 0.5	85.4 ± 0.4*	96.0 ± 0.1*	53.4 ± 0.5*	91.8 ± 0.5	90.4 ± 0.6*	95.2 ± 0.2
$\mathcal{T}_{\not\in Gynecological}$	89.3 ± 0.5	92.4 ± 0.3	86.4 ± 0.2*	96.1 ± 0.3	89.2 ± 0.4	69.1 ± 1.2*	90.8 ± 0.4*	96.2 ± 0.2
$\mathcal{T}_{\not\in Cerebral}$	87.8 ± 0.6	93.7 ± 0.2	85.6 ± 0.6	96.8 ± 0.3*	89.0 ± 0.4	93.0 ± 0.2*	68.8 ± 0.5*	96.7 ± 0.2*
$\mathcal{T}_{\not\in Ophthalmological}$	89.5 ± 1.0	93.1 ± 0.7	86.6 ± 0.9*	96.9 ± 0.4*	89.4 ± 1.0	93.6 ± 0.3*	90.9 ± 0.3*	61.5 ± 2.1*
$\mathcal{T}$	88.9 ± 1.1	93.0 ± 0.8	85.2 ± 0.8	95.4 ± 0.2	87.8 ± 0.8	90.9 ± 0.6	88.7 ± 0.3	94.6 ± 0.7

Table 8: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\in x}$ and $\mathcal{T}$ of the bleeding classification dataset. Means are computed as performance of models trained on $\mathcal{T}_{\not\in x}$ minus $\mathcal{T}$. Total = difference of means on the full test set.

	Total	**Gastrointestinal**	**Urogenital**	**Internal**	**Otorhinolaryngeal**	**Dermatological**	**Gynecological**	**Cerebral**	**Ophthalmological**
	ELECTRA								
$\mathcal{T}_{\not\in Gastrointestinal}$	-1.5 , -0.9	-23.6 , -15.0	0.2 , 2.3	-0.6 , 2.0	0.5 , 1.5	0.6 , 3.3	1.6 , 3.3	1.4 , 2.6	0.5 , 1.1
$\mathcal{T}_{\not\in Urogenital}$	-2.0 , -1.4	0.6 , 2.8	-33.2 , -24.7	1.0 , 2.7	0.9 , 2.1	1.8 , 5.0	1.2 , 3.6	1.8 , 3.8	-0.6 , 1.3
$\mathcal{T}_{\not\in Internal}$	-0.8 , -0.4	0.2 , 2.4	0.7 , 2.4	-18.5 , -9.1	0.1 , 2.2	0.1 , 4.6	0.4 , 3.9	1.2 , 3.0	0.1 , 1.2
$\mathcal{T}_{\not\in Otorhinolaryngeal}$	-0.3 , 0.4	-0.2 , 1.6	0.2 , 1.7	-0.5 , 1.9	-2.9 , -2.1	0.7 , 4.0	0.7 , 2.5	1.5 , 2.6	0.2 , 1.4
$\mathcal{T}_{\not\in Dermatological}$	-1.8 , -1.1	0.0 , 2.4	0.4 , 1.7	0.1 , 1.3	-0.2 , 1.6	-23.5 , -19.8	1.8 , 3.6	1.9 , 2.9	-0.1 , 1.2
$\mathcal{T}_{\not\in Gynecological}$	-2.2 , -1.5	1.1 , 2.3	0.7 , 2.1	0.8 , 2.2	1.3 , 2.7	1.6 , 4.9	-29.5 , -27.3	2.1 , 3.4	0.6 , 1.4
$\mathcal{T}_{\not\in Cerebral}$	-1.5 , -0.5	-0.1 , 1.6	0.2 , 1.9	-0.2 , 2.3	0.5 , 1.6	0.4 , 3.5	1.7 , 3.3	-21.3 , -18.7	0.3 , 1.3
$\mathcal{T}_{\not\in Ophthalmological}$	-0.6 , -0.1	0.5 , 2.3	0.5 , 2.1	0.1 , 1.7	-0.7 , 0.7	-0.1 , 3.7	1.0 , 3.4	1.5 , 2.9	-8.6 , -6.5
	LSTM								
$\mathcal{T}_{\not\in Gastrointestinal}$	-1.9 , -1.4	-30.3 , -18.6	1.0 , 2.1	1.0 , 2.2	0.1 , 1.4	-0.6 , 1.4	-0.2 , 2.6	0.8 , 2.9	1.2 , 3.0
$\mathcal{T}_{\not\in Urogenital}$	-2.1 , -1.8	-1.7 , 4.0	-38.6 , -34.0	-0.5 , 3.6	1.4 , 2.8	0.1 , 4.1	-1.3 , 2.6	1.2 , 3.1	-0.1 , 3.0
$\mathcal{T}_{\not\in Internal}$	-1.8 , -1.4	-4.6 , 0.7	-1.4 , 1.7	-23.0 , -20.3	-1.4 , 1.1	-2.1 , 1.0	-2.6 , 0.2	-2.6 , 0.2	0.0 , 3.0
$\mathcal{T}_{\not\in Otorhinolaryngeal}$	-0.7 , -0.1	-2.7 , 1.8	-0.8 , 2.1	-1.2 , 3.4	-6.9 , -4.5	-1.3 , 4.1	-0.6 , 3.0	0.1 , 2.6	0.2 , 2.7
$\mathcal{T}_{\not\in Dermatological}$	-2.2 , -2.0	-1.1 , 4.2	-2.2 , 2.5	-1.3 , 3.0	0.2 , 1.0	-36.3 , -32.1	-0.2 , 2.0	0.4 , 3.4	-1.0 , 2.2
$\mathcal{T}_{\not\in Gynecological}$	-1.5 , -1.1	-1.9 , 2.7	-2.0 , 0.4	0.3 , 3.8	-0.1 , 1.6	-0.4 , 3.7	-24.3 , -19.0	1.4 , 2.8	0.0 , 2.8
$\mathcal{T}_{\not\in Cerebral}$	-1.7 , -1.2	-3.4 , 0.4	-1.0 , 1.7	-0.3 , 3.0	1.0 , 1.8	-0.6 , 3.4	1.4 , 2.7	-21.0 , -19.0	0.6 , 3.4
$\mathcal{T}_{\not\in Ophthalmological}$	-2.6 , -1.8	-2.9 , 3.8	-1.1 , 1.4	0.1 , 4.8	0.7 , 2.4	-0.2 , 3.4	1.4 , 3.9	1.3 , 3.0	-38.5 , -28.5

Table 9: Bleeding test set sensitivity (%) and standard error on an anatomical location by percentage of subgroup samples in the modified training set. * denotes a significant difference at the 0.05 level between models trained on the modified training set and the full training set, $\mathcal{T}$.

	Anatomical subgroup fraction						
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**	**1.0**
	ELECTRA						
Gastrointestinal	75.7±2.4*	76.6± 0.6*	90.4± 1.1*	93.2± 0.3*	94.5± 0.5*	96.6± 0.5	95.8± 0.4
Urogenital	68.5±2.2*	72.4± 1.1*	91.0± 1.3*	96.6± 0.3*	97.1± 0.1*	97.0± 0.1*	98.4± 0.4
Internal	82.3±2.0*	84.3± 1.8*	91.8± 0.6*	93.6± 0.2*	94.8± 0.5*	95.9± 0.4*	96.4± 0.4
Otorhinolaryngeal	92.6±0.4*	92.2± 0.7*	92.9± 0.3*	94.8± 0.2*	95.4± 0.6	95.8± 0.2	95.6± 0.3
Dermatological	70.6±1.4*	74.6± 1.3*	87.9± 0.7*	88.7± 0.9*	93.3± 0.7	93.4± 0.3	94.5± 0.6
Gynecological	64.8±0.8*	63.4± 1.3*	87.3± 0.6*	89.2±0.3*	92.0± 0.7*	93.7± 0.2	93.9± 0.4
Cerebral	75.4±0.7*	80.0± 1.5*	89.4± 0.4*	91.5± 0.7*	95.6± 0.5*	96.6± 0.6	97.2± 0.3
Ophthalmological	90.70±0.5*	93.6± 0.3*	96.7± 0.4*	97.8± 0.3*	98.8± 0.1	98.7± 0.1	98.9± 0.2
	LSTM						
Gastrointestinal	63.6 ± 3.0*	60.0 ± 1.0*	71.9 ± 1.7*	79.2 ± 0.7*	86.6 ± 0.5*	88.0 ± 0.9*	89.1 ± 0.6
Urogenital	56.7 ± 1.6*	58.6 ± 0.9*	79.3 ± 0.7*	87.2 ± 0.6*	90.6 ± 0.2*	92.2 ± 0.3*	93.7 ± 0.1
Internal	62.6 ± 1.2*	65.6 ± 0.8*	72.6 ± 0.8*	77.2 ± 0.5*	80.2 ± 0.8*	83.9 ± 0.5	85.5 ± 0.5
Otorhinolaryngeal	89.8 ± 0.5*	88.5 ± 0.3*	91.8 ± 0.3*	93.6 ± 0.3*	93.8 ± 0.1*	95.2 ± 0.5	95.5 ± 0.3
Dermatological	53.4 ± 0.5*	56.5 ± 1.5*	73.5 ± 2.4*	77.4 ± 0.6*	84.5 ± 0.3*	86.8 ± 0.4	88.1 ± 0.5
Gynecological	69.1 ± 1.2*	69.3 ± 0.5*	80.1 ± 0.7*	86.6 ± 0.7*	89.3 ± 0.5*	90.7 ± 0.4*	91.6 ± 0.3
Cerebral	68.8 ± 0.5 *	72.2 ± 1.6*	79.7 ± 0.4*	83.1 ± 0.6*	86.4 ± 0.4*	88.4 ± 0.5*	90.5 ± 0.2
Ophthalmological	61.5 ± 2.1*	65.4 ± 1.4*	89.4 ± 0.4*	90.3 ± 0.5*	93.0 ± 0.7*	94.4 ± 0.4*	95.7 ± 0.4

Table 10: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the bleeding classification dataset. Means are computed as performance of models trained on the modified training set minus $\mathcal{T}$.

	Anatomical subgroup fraction					
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**
	ELECTRA					
Gastrointestinal	-23.6 , -15.0	-20.2 , -18.2	-8.3 , -2.6	-4.2 , -1.2	-2.6 , -0.2	-0.2 , 1.8
Urogenital	-33.2 , -24.7	-27.7 , -24.0	-10.0 , -4.8	-2.6 , -1.4	-2.2 , -0.4	-2.2 , -0.6
Internal	-18.5 , -9.1	-15.2 , -9.0	-5.8 , -3.4	-3.8 , -1.8	-2.4 , -1.0	-3.0 , -0.1
Otorhinolaryngeal	-2.9 , -2.1	-4.8 , -2.1	-3.5 , -1.8	-1.5 , -0.2	-0.7 , 0.3	-0.6 , 1.0
Dermatological	-23.5 , -19.8	-22.6 , -16.5	-7.8 , -5.0	-7.8 , -3.4	-3.0 , 1.0	-2.4 , 0.1
Gynecological	-29.5 , -27.3	-33.6 , -27.5	-8.1 , -5.3	-5.5 , -3.8	-3.6 , -0.8	-0.8 , 1.5
Cerebral	-21.3 , -18.7	-20.1 , -13.8	-8.4 , -7.3	-7.0 , -4.4	-2.6 , -0.7	-1.3 , 0.0
Ophthalmological	-8.6 , -6.5	-6.0 , -4.5	-2.7 , -1.8	-1.8 , -0.5	-0.6 , 0.6	-0.2 , 0.6
	LSTM					
Gastrointestinal	-31.7 , -19.8	-31.0 , -27.2	-21.4 , -14.3	-12.3 , -7.6	-4.1 , -0.6	-1.9 , -0.1
Urogenital	-39.8 , -33.8	-36.7 , -32.9	-15.8 , -13.0	-7.8 , -4.8	-3.5 , -2.5	-2.1 , -0.9
Internal	-25.7 , -19.6	-22.3 , -17.4	-15.0 , -10.6	-9.5 , -7.1	-7.4 , -3.1	-3.4 , 0.4
Otorhinolaryngeal	-7.0 , -4.5	-8.2 , -6.2	-4.5 , -3.0	-2.5 , -1.4	-2.3 , -1.1	-1.8 , 1.0
Dermatological	-35.6 , -33.8	-35.5 , -28.6	-19.0 , -10.1	-12.8 , -9.0	-4.9 , -2.5	-3.1 , 0.0
Gynecological	-25.4 , -19.5	-23.3 , -21.2	-13.4 , -10.3	-6.5 , -3.4	-3.9 , -0.9	-1.2 , -0.6
Cerebral	-22.3 , -21.2	-21.1 , -15.1	-11.8 , -9.8	-8.7 , -5.8	-5.0 , -3.2	-3.4 , -0.8
Ophthalmological	-38.4 , -30.5	-33.4 , -27.2	-7.4 , -5.1	-6.5 , -4.1	-3.9 , -1.1	-2.1 , -0.6

Table 11: Precision, recall, and F1 performance for the VTE classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. SE = Standard error. CI = 95% confidence interval.

	ELECTRA			LSTM		
	Precision ± SE (CI)	**Recall ± SE (CI)**	**F1 ± SE (CI)**	**Precision ± SE (CI)**	**Recall ± SE (CI)**	**F1 ± SE (CI)**
$\mathcal{T}_{\not\subset Lower\ extremity}$	86.3 ± 0.7 (84.9 - 87.7)	41.8 ± 1.5 (39.1 - 44.7)	56.3 ± 1.3 (53.8 - 58.8)	77.2 ± 0.2 (76.9 - 77.6)	61.4 ± 1.5 (58.8 - 64.5)	68.3 ± 0.8 (66.8 - 70.1)
$\mathcal{T}_{\not\subset Lung}$	86.1 ± 1.0 (84.1 - 87.9)	57.4 ± 3.2 (51.8 - 63.8)	68.6 ± 2.0 (64.9 - 72.7)	78.6 ± 0.5 (77.6 - 79.6)	54.8 ± 0.7 (53.5 - 56.1)	64.6 ± 0.3 (63.9 - 65.3)
$\mathcal{T}$	96.4 ± 0.3 (95.9 - 97.0)	72.3 ± 0.7 (71.0 - 73.6)	82.7 ± 0.4 (81.9 - 83.5)	91.4 ± 0.1 (91.2 - 91.6)	57.2 ± 0.6 (56.0 - 58.4)	70.4 ± 0.4 (69.4 - 71.2)

Table 12: Sensitivity (%) and standard error for all anatomical locations of the VTE classification dataset. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$.

	Lower extremity	**Lung**	**Liver**	**Cerebral**	**Upper extremity**
			ELECTRA		
$\mathcal{T}_{\not\subset Lower\ extremity}$	9.3 ± 0.8*	98.6 ± 0.2*	50.1 ± 3.6*	23.8 ± 1.9*	25.3 ± 1.4*
$\mathcal{T}_{\not\subset Lung}$	99.7 ± 0.2*	16.6 ± 3.5*	65.7 ± 7.7	13.6 ± 4.1	99.0 ± 0.3*
$\mathcal{T}$	99.1 ± 0.2	97.6 ± 0.3	66.3 ± 1.6	11.5 ± 1.6	96.7 ± 0.3
			LSTM		
$\mathcal{T}_{\not\subset Lower\ extremity}$	34.2 ± 1.5*	92.6 ± 0.5*	76.2 ± 2.2*	52.6 ± 2.2*	47.6 ± 2.8*
$\mathcal{T}_{\not\subset Lung}$	95.3 ± 0.2*	15.7 ± 0.4*	54.2 ± 1.8*	24.7 ± 1.3*	91.4 ± 0.4*
$\mathcal{T}$	93.8 ± 0.6	86.6 ± 0.6	29.8 ± 1.6	7.0 ± 0.5	81.8 ± 1.3

Table 13: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the VTE classification dataset. Means are computed as performance of models trained on $\mathcal{T}_{\not\subset x}$ minus $\mathcal{T}$. Total = difference of means on the full test set.

	Total	**Lower extremity**	**Lung**	**Liver**	**Cerebral**	**Upper extremity**
				ELECTRA		
$\mathcal{T}_{\not\subset Lower\ extremity}$	-18.0 , -16.3	-91.3 , -88.3	0.2 , 1.6	-23.9 , -7.3	8.0 , 16.6	-74.0 , -68.7
$\mathcal{T}_{\not\subset Lung}$	-12.4, -9.2	0.3 , 0.9	-87.2 , -73.7	-12.6 , 11.6	-2.0 , 8.3	1.8 , 2.7
				LSTM		
$\mathcal{T}_{\not\subset Lower\ extremity}$	-5.1 , -3.5	-63.0 , -56.5	4.3 , 7.7	43.8 , 50.0	40.3 , 49.6	-41.9 , -28.2
$\mathcal{T}_{\not\subset Lung}$	-6.6 , -5.4	0.6 , 2.6	-72.5 , -69.5	18.9 , 28.8	14.9 , 19.7	7.5 , 11.9

Table 14: VTE test set sensitivity (%) and standard error on an anatomical location by percentage of subgroup samples in the modified training set. * denotes a significant difference at the 0.05 level between models trained on the modified training set and the full training set, $\mathcal{T}$.

	Anatomical subgroup fraction						
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**	**1.0**
				ELECTRA			
Lower extremity	9.3 ± 0.8*	77.6 ± 3.7*	97.2 ± 0.3*	97.7 ± 0.4	98.6 ± 0.3	98.8 ± 0.2	98.4 ± 0.2
Lung	16.6 ± 3.5*	57.1 ± 8.1*	94.6 ± 0.3*	96.6 ± 0.3	97.7 ± 0.2	98.2 ± 0.1	97.6 ± 0.3
				LSTM			
Lower extremity	34.2 ± 1.5	40.6 ± 0.7	83.7 ± 0.6	90.7 ± 0.5	94.4 ± 0.6	95.5 ± 0.6	94.9 ± 0.7
Lung	15.7 ± 0.4	16.0 ± 0.9	71.5 ± 1.1	81.6 ± 0.7	88.0 ± 0.9	88.3 ± 0.9	89.2 ± 0.8

Table 15: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the VTE classification dataset. Means are computed as performance of models trained on the modified training set minus $\mathcal{T}$.

	Anatomical subgroup fraction					
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**
			ELECTRA			
Lower extremity	-91.3 , -88.3	-28.8 , -15.0	-2.2 , -0.4	-1.9 , 0.4	-0.2 , 0.7	-0.4 , 1.1
Lung	-87.2 , -73.7	-58.8 , -27.4	-3.6 , -2.2	-2.0 , 0.0	-0.7 , 0.8	-0.1 , 1.2
			LSTM			
Lower extremity	-63.0 , -56.5	-56.2 , -52.2	-13.2 , -9.4	-5.6 , -2.9	-2.3 , 1.2	-1.3 , 2.2
Lung	-72.5 , -69.5	-76.3 , -70.1	-20.6 , -14.4	-9.2 , -6.0	-3.5 , 0.8	-3.6 , 1.9

Table 16: Most frequent words used to describe bleeding mentions for each anatomical location of the bleeding classification dataset. Words are translated from Danish to English and, therefore, some cells include two words.

Word	**Frequency**	**Location uniqueness**	**Word**	**Frequency**	**Location uniqueness**
Location: Otorhinolaryngeal			Location: Gynecological		
bleeding	324	0.13	bleeding	714	0.29
nose bleeding	273	1.0	uterus	108	0.97
epistaxis	254	0.99	allowable	78	0.94
nostril	148	1.0	vagina	69	1.0
Location: Dermatological			Location: Cerebral		
haematoma	354	0.56	sah	217	0.99
bleeding	170	0.07	bleeding	190	0.08
skin	122	0.73	ct	185	0.63
right	97	0.29	haematoma	161	0.25
Location: Urogenital			Location: Internal		
haematuria	536	0.99	bleeding	273	0.11
urine	311	0.98	haemothorax	249	1.0
blood	205	0.23	fluid	174	0.80
macroscopic	186	0.99	blood	140	0.16
Location: Gastrointestinal			Location: Ophthalmological		
bleeding	560	0.23	corpus hemorrhagicum	270	1.0
blood	247	0.28	corpus hem	205	1.0
fresh	180	0.48	bleeding	198	0.08
melaena	165	0.98	haemorrhage	180	0.70

CHAPTER 5

Language models for information extraction in Danish electronic health records

This chapter presents our work on utilizing language models for information extraction in the narrative text of EHRs.

Section 5.1 presents the paper *Deep learning detects and visualizes bleeding events in electronic health records* [18] where we presented the first study for extracting bleeding events from Danish EHRs.

Section 5.2 presents our paper *Doctors identify bleedings better during chart review when assisted by artificial intelligence* [19]. In this follow-up paper, we improved the performance of the bleeding detection algorithm, and we presented a user study that investigated MDs' performance when using the algorithm in a clinical scenario and the MDs' perception of using it.

Section 5.3 presents our paper *Danish Clinical Named Entity Recognition and Relation Extraction* [20]. In this paper we presented the first Danish clinical NER dataset and an associated language model that can be used to extract six types of clinical events, six types of attributes, and three types of relations in Danish EHRs.

5.1 Extraction of bleeding events from the narrative text of Danish electronic health records

As written in the introduction, previous bleeding events are an important factor for predicting the risk of future bleeding and this information is therefore used in clinical guidelines for administering prophylactic medicine. In the paper *Deep learning detects and visualizes bleeding events in electronic health records* [18] we presented our work on developing a dataset with EHR paragraphs annotated as positive or negative for bleeding mentions. The dataset was used to

develop NLP algorithms for extracting bleeding descriptions from Danish EHRs. As this study was conducted prior to the development of the Clin-ELECTRA model, we used other deep learning-based architectures with clinical word embeddings to develop the automatic bleeding detection algorithm.

We found that a hybrid RNN-CNN-based architecture could classify and extract bleeding events in the free text of EHRs with an accuracy of 90%. Moreover, we argued that automatic bleeding detection algorithms can be used to make a fast overview of a patient's bleeding history by highlighting the relevant text in the narrative part of the EHR. However, the performance of the classifier should be improved before being taken into use in clinical practice.

The paper is presented in its original form below. It is licensed under a CC BY-NC-ND 4.0 license which can be found at `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

Received: 25 January 2021 | Revised: 21 February 2021 | Accepted: 2 March 2021

DOI: 10.1002/rth2.12505

ORIGINAL ARTICLE

Deep learning detects and visualizes bleeding events in electronic health records

Jannik S. Pedersen MSc Eng[1] | Martin S. Laursen MSc Eng[1] | Thiusius Rajeeth Savarimuthu PhD[1] | Rasmus Søgaard Hansen MD[2] | Anne Bryde Alnor MD[2] | Kristian Voss Bjerre MD[2] | Ina Mathilde Kjær MD[3] | Charlotte Gils MD[2] | Anne-Sofie Faarvang Thorsen MD[2] | Eline Sandvig Andersen MD[3] | Cathrine Brødsgaard Nielsen MD[4] | Lou-Ann Christensen Andersen MD[5] | Søren Andreas Just MD[6] | Pernille Just Vinholt MD[2]

[1]The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark

[2]Department of Clinical Biochemistry and Pharmacology, Odense University Hospital, Odense, Denmark

[3]Department of Clinical Biochemistry and Immunology, Lillebaelt Hospital, Denmark

[4]Department of Gastroenterology, Odense University Hospital, Odense, Denmark

[5]Department of Ophthalmology, Odense University Hospital, Odense, Denmark

[6]Department of Medicine, Odense University Hospital, Svendborg, Denmark

Correspondence
Jannik S. Pedersen, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Campusvej 55, Odense, Denmark.
Email: jasp@mmmi.sdu.dk

Handling Editor: Dr Lana Castellucci

Abstract

Background: Bleeding is associated with a significantly increased morbidity and mortality. Bleeding events are often described in the unstructured text of electronic health records, which makes them difficult to identify by manual inspection.

Objectives: To develop a deep learning model that detects and visualizes bleeding events in electronic health records.

Patients/Methods: Three hundred electronic health records with *International Classification of Diseases*, *Tenth Revision* diagnosis codes for bleeding or leukemia were extracted. Each sentence in the electronic health record was annotated as positive or negative for bleeding. The annotated sentences were used to develop a deep learning model that detects bleeding at sentence and note level.

Results: On a balanced test set of 1178 sentences, the best-performing deep learning model achieved a sensitivity of 0.90, specificity of 0.90, and negative predictive value of 0.90. On a test set consisting of 700 notes, of which 49 were positive for bleeding, the model achieved a note-level sensitivity of 1.00, specificity of 0.52, and negative predictive value of 1.00. By using a sentence-level model on a note level, the model can explain its predictions by visualizing the exact sentence in a note that contains information regarding bleeding. Moreover, we found that the model performed consistently well across different types of bleedings.

Conclusions: A deep learning model can be used to detect and visualize bleeding events in the free text of electronic health records. The deep learning model can thus facilitate systematic assessment of bleeding risk, and thereby optimize patient care and safety.

Jannik S. Pedersen and Martin S. Laursen are co-first authors.

rpth research & practice in thrombosis & haemostasis

KEYWORDS
decision support systems (clinical), deep learning, electronic health record, hemorrhage, international classification of diseases, machine learning

Essentials

- Bleeding events are difficult to locate in electronic health records.
- A deep learning model detects bleeding events and visualizes them to the clinicians.
- The model identified 90.0% of bleeding-positive sentences and 89.6% of negative sentences
- The model identified 100% of bleeding-positive notes and 52.4% of negative notes.

1 | INTRODUCTION

Bleeding occurs in 3.2% of medical patients within 14 days of admission, and approximately one-third of the bleeding events are considered major events.[1] Bleeding is associated with a significantly increased morbidity and mortality.[2,3] Furthermore, previous clinically relevant bleeding events are a strong independent risk factor for future bleeding.[1] Hence, knowledge about bleeding history is essential for providing optimal care to patients.

In clinical practice, bleeding risk can be assessed using bleeding risk scores that include information about the patient's bleeding history, for example the HAS-BLED (hypertension, abnormal renal and liver function, stroke, bleeding, labile international normalized ratio, elderly, drugs or alcohol) score, which is recommended for determining bleeding risk during anticoagulation treatment,[4,5] or the IMPROVE (International Medical Prevention Registry on Venous Thromboembolism) score, which is recommended to guide prophylactic anticoagulant treatment for adult medical patients at admission.[1]

Although crucial for patient care, bleeding risk is not always systematically evaluated. Studies have shown that a large proportion of hospitalized medical patients do not get appropriate prophylactic anticoagulant treatment during admission.[6-8] One reason is that the recommended scoring systems for assessment of thrombosis and bleeding risk are not always used in clinical practice.[6-8] This could be caused by the fact that risk scores are laborious to obtain because it requires manual work to go through the electronic health record (EHR) for relevant information[7] and that it must be done at the time of admission when health care professionals are busy handling the acute situation.

In recent years, deep learning techniques have achieved state-of-the-art performance on text classification benchmarks.[9] In medicine, various deep learning techniques have been used for text classification including, but not limited to, recurrent neural networks (RNNs),[10-12] convolutional neural networks (CNNs),[13,14] and hybrid models combining more than one technique.[15] These techniques have the potential for automatic detection of relevant clinical information in EHR text. This could facilitate the systematic assessment of bleeding risk and thereby optimize patient care and safety as well as freeing up time for health care professionals. To date, only a few studies have used deep learning for finding bleeding events in EHRs.[15-17]

A general concern about deep learning is how the models reach their conclusions. It often remains a black box, making the users struggle to assess the basis for results or whether the model answers the questions for which clinicians want assistance.[18,19] Therefore, there is a growing awareness that deep learning models need to be self-explanatory.[20] For text classification models, it means that it is relevant to show the prediction-supporting part of the text upon request. However, such approaches are lacking in bleeding detection models.

Therefore, the purpose of this study was to establish a deep learning model that automatically detects bleeding events on a sentence level and to visualize the bleeding events to the clinician in the unstructured EHR text.

2 | METHODS

2.1 | Population and data set

Data were acquired from the EHR system of the Region of Southern Denmark. To ensure inclusion of EHR notes with a high likelihood of bleeding events in the text, we extracted EHRs from 300 patients with *International Classification of Diseases, Tenth Revision* (*ICD-10*) diagnosis codes for bleeding or leukemia. *ICD-10* codes for bleeding from the following sites were included: eyes, ear-nose-throat and respiratory tract, gastrointestinal, urogenital, internal organs, hematoma, and others. EHRs from patients with leukemia were included, as this patient group has a high incidence of bleeding (see Appendix S1 for *ICD-10* codes).[21] Before annotation, we discarded administrative notes, as they would not contain any bleeding events.

Twelve physicians annotated the 300 EHRs. Each EHR was annotated by one physician. To determine the agreement between physicians' annotation, we calculated the kappa score on a sample of 1328 sentences from randomly chosen EHRs.

The EHRs were annotated[22] on sentence level with two different labels:

1. Positive: Sentences that indicate any kind of bleeding.
2. Misinterpretable negative: Sentences that were deemed by the annotator to have a high risk of being misinterpreted by the deep learning model, for example, "The patient is not bleeding."

rpth research & practice in thrombosis & haemostasis

All sentences left after annotation of positive and misinterpretable negative sentences were then considered negative sentences. We chose to annotate the misinterpretable negative sentences as a subcategory to the negative category to be able to feed many negative samples that resemble positive samples to the model. This should help the model distinguish for example "the patient has a bleeding" from "the patient might have a bleeding."

Data were split into a balanced training (80%), validation (10%), and test set (10%) using subsampling of the overrepresented class.[23] The negative sentences consisted of 50% random negatives and 50% misinterpretable negatives. The training set was used to train the models, the validation set was used to tune parameters of the models during training, and the test set was used to evaluate final performance.

Sentences were tokenized using the Stanza sentence tokenizer.[24] Samples were preprocessed by elimination of superfluous spaces, special characters, and duplicate sentences.

2.2 | Models for detection of bleeding events on sentence level

2.2.1 | Rule-based classifier

A rule-based classifier was developed to compare the deep learning models with a traditional approach to text classification.

The rule-based model was constructed by defining a set of bleeding-indicating words and modifiers using corpus statistics and manual inspection of the data. Corpus statistics were used to calculate the most frequent words in bleeding-indicating sentences. A bleeding-indicating word could for example be *bleeding*, and a modifier could, for example, be *no* (bleeding). Next, by evaluating performance on the training data, a window size was defined where a modifier could modify a bleeding-indicating word. For example, *no* would modify *bleeding* in "no sign of bleeding" for a window size of 3. The model uses the indicating and modifying words and the window size to create rules for classifying individual sentences. The rules were iteratively updated during training to improve performance.

2.2.2 | Deep learning models

Three different deep learning models were developed: a CNN model, an RNN model, and a hybrid model combining an RNN and a CNN. In deep learning, a model transforms the input to a classification via many layers of processing steps that are learned from labeled data during training. The input to the models is the individual words from each sentence represented as word embeddings. Word embeddings are numerical vector representations of words that encode their meaning with similar words having similar vectors. For this study, 100-dimensional GloVe word embeddings pretrained on 323 122 Danish EHRs were used.[25,26]

2.3 | Evaluation of internal validity

We performed an internal sensitivity analysis on the best-performing model to evaluate if it performs equally well on the seven patient groups included in the study.

2.4 | Bleeding detection on note level

Because each note may contain multiple positive sentences that often describe the same bleeding event, we calculated the performance of the best model on a note level by classifying all sentences of each note. A positive note is defined as a note that includes at least one bleeding-positive sentence. The test was performed on seven randomly selected EHRs from patients in the leukemia group not included in the original data set. A total of 100 notes per EHR were collected.

2.5 | Visualization of bleeding events in EHR text

Finally, we present how the bleeding-positive output of the model can be presented to the physician as a visualization of complete notes with the bleeding events highlighted, helping the physician understand the prediction and decreasing the time needed to find a bleeding event in an EHR.

2.6 | Statistical analysis

We calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and a harmonic mean of sensitivity and positive predictive value (F1) score. For each model, we plotted receiver operating characteristic curves and calculated area under the receiver operating characteristic curve (AUC).

The models were developed in Python 3.6 (Python Software Foundation, Wilmington, DE, USA) using the Tensorflow 2.0 framework.

3 | RESULTS

The 300 extracted EHRs contained 88 477 notes. Of those, we filtered out 43 602 as administrative notes. The remaining 44 875 EHR notes were annotated on a sentence level. In total, 6111 sentences were annotated as positive and 5630 as misinterpretable negative. Overall, 3973 notes contained bleeding events and there were 1 to 19 positive sentences per note. The EHRs contained 0 to 108 notes with bleeding per patient.

Among the different patient groups, "gastrointestinal bleeding" had the highest average number of positive sentences per EHR (n = 25) while "Hematomas and other bleedings" had the lowest (n = 8; see Table 1). Although the EHRs were extracted on the basis

of *ICD-10* codes for bleeding, 13 EHRs did not contain any information about bleeding ("internal bleedings," n = 2; "eyes," n = 5; and "hematomas and others," n = 6). Another 5 EHRs from leukemia patients did not contain bleeding events.

When assessing agreement among the 12 physicians, they achieved a kappa score of 0.75 on a sample of 1328 sentences from randomly chosen EHRs. This is considered a substantial agreement.[27]

3.1 | Establishing models

For development of models, we removed duplicate sentences (n = 218), resulting in 5893 positive samples. To create a balanced data set, we randomly subsampled 2947 misinterpretable negative sentences. These were added to 2946 randomly extracted negative samples to give 5893 total negative samples. Together with the 5893 positive samples, they constitute the balanced data set of 11 786 samples.

The balanced data set was divided into training (n = 9430), validation (n = 1178), and test sets (n = 1178). The distribution is seen in Table S1.

3.1.1 | Rule-based results

The bleeding-indicating and modifying words were aggregated into a stem to capture different conjugations; for example, the Danish word for *hemorrhage* (*hæmoragi*) was aggregated to *hæm* and defined as a bleeding-indicating word.

The developed rule-based classification model searched each sentence for a positive word. If no positive words were found, the sentence was classified as negative. If a positive word was found, the model searched its context words in a window of size 4 to look for negative modifiers. If a word from the positive list was not accompanied by a negative modifier, the sample was classified as positive. If all positive words were accompanied by negative modifiers, the sample was classified as negative.

3.1.2 | Deep learning

The developed CNN consisted of convolutional layers that extract information from neighboring words. The extracted information was used by a linear classification layer that classifies the sentence as either bleeding present or bleeding absent.

Our RNN model was based on the Bidirectional Gated Recurrent Unit (BiGRU).[28] The model consisted of a single BiGRU layer that extracts information from the input words by processing them sequentially. The extracted information was used by a linear classification layer that classifies the sentence.

The hybrid model used the output from both a CNN and an RNN to classify the sentences. This model was developed to exploit the information extracted from both a CNN and RNN in a final linear classification layer.

A more thorough description of the models can be seen in Appendix S2.

For each deep learning model, the seven versions of the model that performed best on the validation set were selected for an ensemble classifier. The ensemble classifier averages the predictions of each model to a final prediction.

3.2 | Performance of models for bleeding detection in EHRs on sentence level

Table 2 shows the performance of the rule-based and deep learning classifiers on the test set.

Figure 1 shows the ROC curves of the hybrid, CNN, RNN, and rule-based models with their corresponding AUC.

Overall, the performance of the hybrid model was the best. It achieved an F1 score of 0.90, a sensitivity of 0.90, a specificity of 0.90, a PPV of 0.90, and an NPV of 0.90. The CNN model achieved equally high sensitivity of 0.90 but performed slightly worse on the additional metrics, while the RNN performed consistently worse on all metrics against both the hybrid and CNN model. The rule-based model performed worse than all deep learning models.

TABLE 1 Patient group distribution of extracted EHRs

Patient group	Number of EHRs	Number of EHR notes	Number of positive EHR notes	Number of positive sentences	Average number of positive sentences per EHR
Eye bleeding	65	7781	771	1546	24
Ear-nose-throat and respiratory tract bleeding	23	3702	372	532	23
Gastrointestinal bleeding	51	6968	1,055	1,250	25
Urogenital bleeding	45	4409	499	855	19
Internal organ bleeding	45	6078	753	1,082	24
Hematoma and other bleeding	38	5597	229	319	8
Leukemia bleeding	33	10 340	294	527	16
Total	300	44 875	3973	6111	...

Abbreviation: EHR, electronic health record.

rpth research & practice in thrombosis & haemostasis

3.3 | Test of internal validity

We evaluated the hybrid model's performance within each of the different patient groups on sentence level (Figure 2). It was calculated on the full data set including training, validation, and test sets. The model shows an almost equal performance for all patient groups, highest for "eyes" at 0.98, and lowest for "leukemia" at 0.95.

3.4 | Performance of the hybrid model on note level

We further tested the performance of the hybrid model on a note level by classifying all sentences and aggregating the result to the full note. The seven EHRs contained 700 notes, of which 49 were positive. The hybrid model achieved a sensitivity of 1.00, a specificity of 0.52, a PPV of 0.14, an NPV of 1.00, an F1 score of 0.24, and an AUC of 0.76.

TABLE 2 Performance of models for detecting bleeding in electronic health records on sentence level

	Rule-based	CNN	RNN	Hybrid
Accuracy	0.80	0.89	0.89	0.90
Sensitivity	0.86	0.90	0.89	0.90
Specificity	0.72	0.89	0.88	0.90
Positive predictive value	0.76	0.89	0.88	0.90
Negative predictive value	0.84	0.90	0.90	0.90
F1 score	0.81	0.89	0.89	0.90
AUC	0.79	0.89	0.89	0.90

Abbreviations: AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; F1, harmonic mean of sensitivity and positive predictive value; RNN, recurrent neural network.

3.5 | Visualization of bleeding events in EHR text

In this study, we chose to use the sentence-level model on a note level because it makes the model capable of explaining its predictions. The model outputs all notes with predicted bleeding events, highlighting the sentence(s) found to indicate bleeding (representative example translated to English in Figure 3, original in Figure S1).

4 | DISCUSSION

We present a deep learning model that automatically detects bleeding events in EHRs with a sensitivity of 0.90 on sentence level and 1.00 on note level. This enables clinicians to receive automatic visualization of EHR notes with bleeding events. The hybrid model, combining an RNN and a CNN, performed best for bleeding detection on sentence level (F1 = 0.90).

In congruence with our study, others have found that machine learning can be used for finding bleeding in EHRs. Rumeng et al.[15] used a deep learning model to detect bleeding events in sentences of EHRs (F1 = 0.94). The study comprised a data set of 2902 sentences extracted from 878 notes from patients with cardiovascular events. Taggart et al.[17] detected bleeding events at a note level with

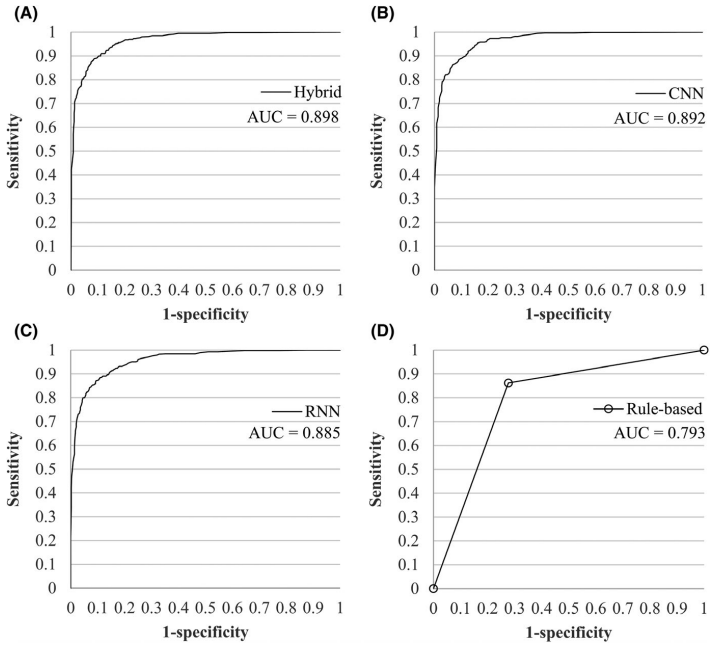


FIGURE 1 ROC curves and AUC for all models on sentence level. (A) Hybrid model. (B) CNN model. (C) RNN model. (D) Rule-based model. AUC, area under the curve; CNN, convolutional neural network; RNN, recurrent neural network; ROC, receiver operating characteristic

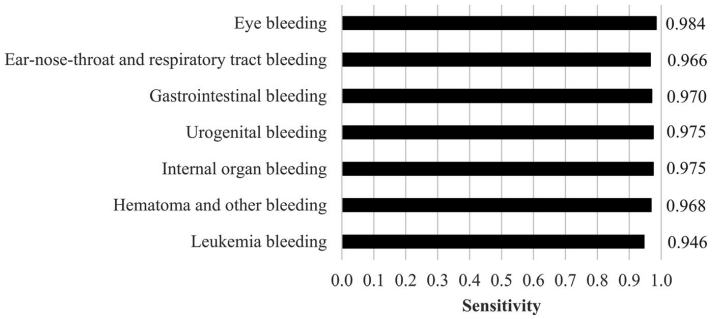


FIGURE 2 Internal validity for detection of bleeding on note level for the hybrid model

Body, Clinical contact. Anamnesis. Hematological anamnesis. Medical secretary []. The patient has no longer diarrhea CRP is lowered to under 10 thrombocytes was recently 8 today 24. The patient is paused with low molecular heparin which the patient is given as prophylaxis partly because of atrial fibrillation and possibly also because of former portal vein thrombosis. The patient has also noticeable bleeding tendency with the low thrombocytes and there has previously been heavy bleeding from esophagus and from esophageal varices. The family explains that a meeting is scheduled at [] to see if clips the spleen should be clipped. I have contacted the physician on call for them to read through the journal and evaluate if this conversation should be maintained or postponed to later. Aim for discharge on []. There has been an increasing need for insulin and we will contact the endocrinologists for advice about this.

FIGURE 3 Example of the visualization of bleeding events in an electronic health record note. To keep the original format, the text is translated directly from Danish to English, which results in incorrect sentence structures

a rule-based approach (F1 = 0.74) and a CNN (F1 = 0.40) on a test set consisting of 660 notes. The rule-based model was trained on 990 notes and the CNN was trained on 450 notes.

In contrast to our study, Taggart et al. found that their rule-based approach performed better than their CNN but it may, however, be due to the limited amount of training data for the CNN, which is a well-known limitation in machine learning.[29,30] Rumeng et al. also used a small data set, and moreover, the data used were exclusively from patients with cardiovascular events. Therefore, in the above studies, the data sets might not be representative of bleeding in all sites and the model might not be generalizable to other patient groups. The model presented in our study used a data set of 11 786 sentences extracted from 44 875 notes representative for multiple types of bleeding. In the internal validity test, we found that our model generalizes well to different types of bleeding.

Lee et al.[16] used a rule-based (sensitivity = 0.83), machine learning (sensitivity = 1.00), and score function (sensitivity = 0.98) approach to find clopidogrel-induced bleedings in EHRs. They defined bleeding events as the presence of specific *ICD*, *Ninth Revision* (*ICD-9*) codes, specific keywords, and unique identifiers of the Unified Medical Language System related to bleeding. Thus, the bleeding definition was simplified to specific words, which is a limitation for use in clinical practice, as bleeding can be reported with numerous different phrases in EHRs. In agreement, we found thousands of different sentences that corresponded to bleeding according to the physicians involved. Moreover, the construction of keyword and rule lists requires manual effort that is difficult to scale because of the unstructured and noisy nature of the clinical notes (eg, grammatical ambiguity, synonyms, term abbreviation, misspelling, or negation of concepts).[31]

Additionally, validation of *ICD-9* and *ICD-10* diagnosis codes has shown that they are not always accurate.[32,33] However, the major concern is that diagnosis coding requires manual collection of the patient history to choose the codes of relevance and that bleeding events that are not a major contributing cause of admittance are not registered with a code for bleeding. The present study provides an attractive alternative by leveraging the information-rich yet unstructured text data in clinical notes in EHRs, which are currently often omitted when developing models.[34]

In the present approach, we established a deep learning model that points out relevant information in the EHRs on sentence level. The advantage of making a sentence-level classifier is that it enables the model to explain its predictions on a note level by showing the prediction-supporting part of the text. We were, therefore, able to visualize where in the notes the model has detected a bleeding event, which enables us to point out relevant sentence(s) in the long unstructured EHR text for the physician. A fast overview of patient bleeding history facilitates clinical decision making. Accordingly, studies have shown that clinical practice may improve when decision support systems give automatic recommendations where the decision is interpretable and understandable for the physician.[35,36]

An automatic summary of bleeding history may be valuable in clinical practice to diagnose, monitor disease, or address treatment options. The presented approach can be extended to include other symptoms and findings. Information regarding specific past events, for example, bleeding events during medical procedures, is

important when planning a new medical procedure. Thus, the information may have an impact on patient safety because, for example, procedure and operation bleeding risk and medication side effects can be monitored effectively. It may also prove useful for health care statistics and resource management. Finally, the approach may save time because a focused review of an EHR to find all past bleeding events is very time consuming. Thus, it provides more time for direct patient care.

To summarize the main points of the discussion, comparing the related studies, the current study used the largest annotated corpus, providing an advantage to the deep learning model. This study also included many different types of bleeding, and it evaluated model accuracy by type of bleeding. In contrast to Taggart et al., we found that a deep learning approach works better than a rule-based approach. We additionally show a simple approach to visualizing the sentences indicating bleeding to physicians, allowing for interpretation of the deep learning model.

4.1 | Limitations

The rule-based algorithm may have been further optimized by being more specific on search terms with inclusion of more words and their common misspellings instead of using more global terms to group words; for example, the Danish stem *hæm* may find words with various meanings that do not imply bleeding. Another limitation is that the study included only EHRs with an *ICD-10* code of bleeding, which does not capture all EHRs with bleeding events. Additionally, we did not validate the algorithm on an independent cohort. Of note, we found a high sensitivity for bleeding in EHRs from patients with leukemia, who comprise a patient group experiencing bleeding from different organ systems.[21] It thus suggests that the model performs well on EHRs without *ICD-10* for bleeding. It is crucial that the text that we used for training the model is representative for any way that bleeding can be reported in the EHR. It is a limitation to the study that we cannot guarantee this, and it would be beneficial to include a larger and more general data set. Nevertheless, this approach clearly showed that it is feasible to automatically extract and visualize bleeding events in EHRs. Future research shall focus on developing a model on data including even more bleeding types and optimizing the strategy, which includes differentiation between clinically relevant versus trivial bleedings and surgical versus medical bleeding.

5 | CONCLUSION

We have developed a deep learning model that identifies bleeding events in EHRs with a sensitivity of 0.90 on sentence level and 1.00 on note level. Further, we have shown how bleeding-positive notes can be visualized to physicians, making the model easily interpretable to the clinician.

ACKNOWLEDGMENTS

The authors thank Helene Kirkegaard, Camilla Brødsgaard Nielsen, and Kristina Bjerg Appel for contributing to the annotation process, as well as Jan Hellden for extracting data.

AUTHOR CONTRIBUTIONS

MSL and JSP contributed equally in the analysis of data and production of results. PJV, TRS, and SAJ contributed to the design and conception of the study. PJV, RSH, ABA, KVB, IMK, CG, AFT, ESA, CBN, and LCA annotated the electronic health records.

RELATIONSHIP DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Decousus H, Tapson VF, Bergmann JF et al. Factors at admission associated with bleeding risk in medical patients: findings from the IMPROVE investigators. *Chest*. 2011;139(1):69-79. https://doi.org/10.1378/chest.09-3081
2. Berger JS, Bhatt DL, Steg PG et al. Bleeding, mortality, and antiplatelet therapy: Results from the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial. *Am Heart J*. 2011;162(1):98-105.e1. https://doi.org/10.1016/j.ahj.2011.04.015.
3. Cook DJ, Griffith LE, Walter SD et al. The attribute mortality and length of intensive care unit stay of clinically important gastrointestinal bleeding in critically ill patients. *Crit Care*. 2001;5(6):368-375. https://doi.org/10.1186/cc1071.
4. Pisters R, Lane DA, Nieuwlaat R et al. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: The Euro Heart Survey. *Chest*. 2010;138(5):1093-1100. https://doi.org/10.1378/chest.10-0134.
5. Rodeghiero F, Tosetto A, Abshire T et al. ISTH/SSC bleeding assessment tool: a standardized questionnaire and a proposal for a new bleeding score for inherited bleeding disorders. *J Thromb Haemost*. 2010;8(9):2063-2065. https://doi.org/10.1111/j.1538-7836.2010.03975.x.
6. Amin A, Stemkowski S, Lin J, Yang G. Thromboprophylaxis rates in US medical centers: success or failure? *J Thromb Haemost*. 2007;5(8):1610-1616. https://doi.org/10.1111/j.1538-7836.2007.02650.x.
7. Rwabihama JP, Audureau E, Laurent M et al. Prophylaxis of venous thromboembolism in geriatric settings: a cluster-randomized multicomponent interventional trial. *J Am Med Dir Assoc*. 2018;19(6):497-503. https://doi.org/10.1016/j.jamda.2018.02.004.
8. Amin A, Stemkowski S, Lin J, Yang G. Appropriate thromboprophylaxis in hospitalized cancer patients. *Clin Adv Hematol Oncol*. 2008;6(12):910-920.
9. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning based text classification: a comprehensive review. arXiv Prepr. Published online 2020. https://arxiv.org/pdf/2004.03705.pdf#:~:text=Deeplearningbasedmodelshave,answering%2Candnaturallanguageinference
10. Santiso S, Pérez A, Casillas A. Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery. *IEEE J Biomed Heal Informatics*. 2019;23(5):2148-2155. https://doi.org/10.1109/JBHI.2018.2879744.
11. Chen D, Qian G, Pan Q. Breast cancer classification with electronic medical records using hierarchical attention bidirectional networks. In: Proceedings - 2018 IEEE International Conference

on Bioinformatics and Biomedicine, BIBM 2018; 2019. https://doi.org/10.1109/BIBM.2018.8621479
12. Deng Y, Dolog P, Gass JM, Denecke K. Obesity entity extraction from real outpatient records: When learning-based methods meet small imbalanced medical data sets. In: Proceedings - IEEE Symposium on Computer-Based Medical Systems. 2019. https://doi.org/10.1109/CBMS.2019.00087
13. Rajput K, Chetty G, Davey R. Deep neural models for chronic disease status detection in free text clinical records. In: IEEE International Conference on Data Mining Workshops, ICDMW. 2019. https://doi.org/10.1109/ICDMW.2018.00127
14. Hughes M, Li I, Kotoulas S, Suzumura T. Medical text classification using convolutional neural networks. *Stud Health Technol Inform*. 2017;235:246-250. https://doi.org/10.3233/978-1-61499-753-5-246.
15. Li R, Hu B, Liu F et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. *J Med Internet Res*. 2019;7(1):e10788. https://doi.org/10.2196/10788.
16. Lee H-J, Jiang M, Wu Y et al. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci. Published online 2017.
17. Taggart M, Chapman WW, Steinberg BA et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA Netw Open*. 2018;1(6):e183451. https://doi.org/10.1001/jamanetworkopen.2018.3451.
18. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018;320(21):2199. https://doi.org/10.1001/jama.2018.17163.
19. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med*. 2020;172(1):59. https://doi.org/10.7326/M19-2548.
20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. https://doi.org/10.1038/s42256-019-0048-x.
21. Webert KE, Cook RJ, Sigouin CS, Rebulla P, Heddle NM. The risk of bleeding in thrombocytopenic patients with acute myeloid leukemia. *Haematologica*. 2006;91(11):1530-1537.
22. Pustejovsky J, Stubbs A. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc.; 2012.
23. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*. 2008;21(2-3):427-436. https://doi.org/10.1016/j.neunet.2007.12.031.
24. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a python natural language processing toolkit for many human languages. In; 2020. https://doi.org/10.18653/v1/2020.acl-demos.14
25. Rasmussen M, Berggrein N. Named entity recognition and disambiguation in Danish electronic health records. Published online. 2019. https://www.derczynski.com/itu/docs/Named_Entity_Recognition_and_Disambiguation_MSc_Thesis%202019.pdf, Accesed 4 april 2021
26. Pantazos K, Lauesen S, Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. *Health Informatics J*. 2017;23(4):291-303. https://doi.org/10.1177/1460458216647760.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159. https://doi.org/10.2307/2529310.
28. Cho K, Van Merriënboer B, Gulcehre C et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2014. https://doi.org/10.3115/v1/d14-1179
29. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. https://doi.org/10.1109/ICCV.2017.97
30. Banko M, Brill E. Mitigating the paucity-of-data problem: exploring the effect of training corpus size on classifier performance for natural language processing. *Comput Linguist*. 2001. https://doi.org/10.3115/1072133.1072204.
31. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Trans Comput Biol Bioinforma*. 2019;16(1):139-153. https://doi.org/10.1109/TCBB.2018.2849968.
32. Valkhoff VE, Coloma PM, Masclee GMC et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *J Clin Epidemiol*. 2014;67(8):921-931. https://doi.org/10.1016/j.jclinepi.2014.02.020
33. Øie LR, Madsbu MA, Giannadakis C et al. Validation of intracranial hemorrhage in the Norwegian patient registry. *Brain Behav*. 2018;8(2):e00900. https://doi.org/10.1002/brb3.900.
34. Esteva A, Robicquet A, Ramsundar B et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. https://doi.org/10.1038/s41591-018-0316-z.
35. Berner ES, La Lande TJ. (2017). Overview of clinical decision support systems. In *Clinical decision support systems*. (pp. 3-22). Springer, New York, NY.. https://doi.org/10.1007/978-3-319-31913-1_1
36. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Br Med J*. 2005;330(7494):765. https://doi.org/10.1136/bmj.38398.500764.8F

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Pedersen JS, Laursen MS, Rajeeth Savarimuthu T, et al. Deep learning detects and visualizes bleeding events in electronic health records. *Res Pract Thromb Haemost*. 2021;5:e12505. https://doi.org/10.1002/rth2.12505

Supplementals (Online only)

Online only elements:

eText 1: ICD-10 codes defining bleeding types
eText 2: Description of deep learning models
eFigure 1. Original Danish example of the visualization of bleeding events in an electronic health record note.
eFigure 2. Overview of the convolutional neural network architecture.
eFigure 3. Overview of the recurrent neural network architecture.
eFigure 4. Overview of the hybrid architecture combining a convolutional and recurrent neural network.
eTable 1. Train, validation, and test distribution of the sentence level dataset.

eText 1: ICD-10 codes defining bleeding types

Eye bleeding: DH052, DH113, DH210, DH180A, DH313, DH356, DH431, DH448 , DH450, DB303C
Ear-nose-throat and respiratory tract bleeding: DR04, DR040, DR041, DR042, DR048, DR049, DS003A, DH922
Gastrointestinal bleeding: DK250, DK251, DK252, DK254, DK256, DK260, DK262, DK270, DK272, DK274, DK276, DK280, DK282, DK284, DK286, DK290, DK625, DK638B, DK638C, DK766B, DK920, DK921, DI841C, DI844C, DI848C, DI850, DI859, DI864A, DK228F, DK922
Urogenital bleeding: DN02, DN508X, DN308B, DR31
Internal organ bleeding: DK661, DM250, DK838F, DK762, DK858B, DK868G, DS260, DT144A, DE078B, DE274B, DG361, DI230, DI288A, DI312, DN830A, DN831B, DN836, DN837, DN838D, DN838E, DN838F, DN857, DN897, DJ942, DJ942A, DJ950C, DK089A, DG951A
Hematoma and other bleeding: DS601A, DS902A, DT140D, DT140K, DR233, DR233A, DN908D, DD500, DD62, DN488D, DN488E, DN501A, DN501B, DN501C, DN501D, DN501E, DN421
Leukemia bleeding: C91, C92, C93, C94, C95

eText 2: Description of deep learning models

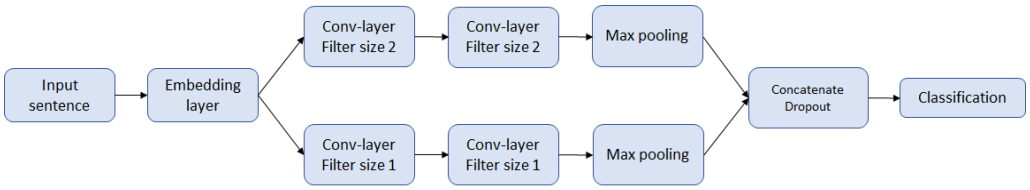
A convolutional neural network (CNN) uses a set of learned filters to recognize local patterns in input data, for example by looking at two words at a time. The advantage of a CNN is its ability to extract patterns from relationships between neighboring words, whereas it is not capable of capturing patterns from words that are far apart.
The developed CNN consists of two parallel series of convolutional layers (conv-layers). The first series has two 1D conv-layers followed by a max-pooling layer. These conv-layers have a filter size of 1 and each 1792 filters. This means that the first series is analyzing the sentences one word at a time with 3584 filters that look for different patterns. The second series has the same structure except that the conv-layers have a filter size of 2, meaning that the sentences are analyzed looking at two words at a time. The outputs from the parallel convolutional series are concatenated and sent to a dense classification layer with a dropout layer in-between (eFigure 2).
A recurrent neural network (RNN) is another kind of neural network that in contrast to CNNs is capable of modelling long term dependencies. The RNN infers meaning from text by processing the input words sequentially, and in this way its final prediction is based on the information it has learned from all the previous input words.
The developed RNN model is based on a type of RNN called Gated Recurrent Unit (GRU). This type of RNN has specialized units that make it capable of forgetting insignificant information and remembering important information. The developed GRU model has a hidden layer size of 16 (eFigure 3).
The hybrid model consists of two parallel layers, one being a CNN and the other a RNN. Both the CNN and RNN are constructed as described above, except that their output representations are concatenated before being sent to the classification layer. This was done to combine the local pattern strengths of the CNN with the capability of modelling long term dependencies of the RNN in a single model (eFigure 4).

---

eFigure 1. Original Danish example of the visualization of bleeding events in an electronic health record note.

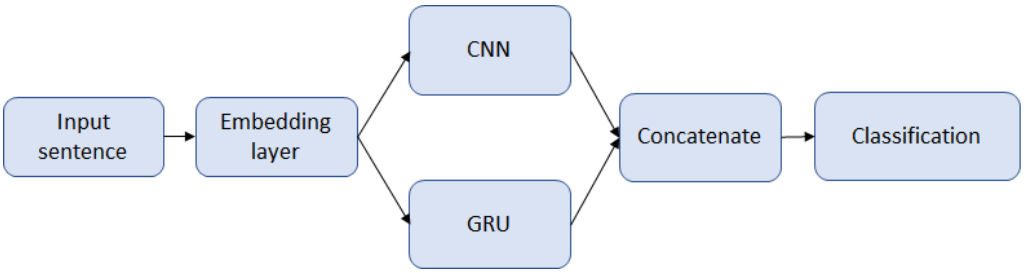
Broedtekst Klinisk kontakt. Anamnese. Hæmatologisk anamnese. Lægesekretær []. Patienten har ikke længere diarré CRP er faldet til under 10 trombocytter var forleden 8 i dag 24. Patienten pauserer med lavmolekylær heparin som patienten får profylaktisk dels på grund af atrieflimmer muligvis også på grund af tidligere porta venetrombose. Patienten har også udtalt blødningstendens med de lave trombocytter, og der har tidligere været voldsom blødning fra øsofagus og fra fundus varicer. Familien fortæller i dag, at der er planlagt en samtale på [] med henblik på, om der skulle clipses på milten. Jeg har kontaktet bagvagt med henblik på, om de vil læge journalen igennem og vurdere, om denne samtale skal fastholdes eller udskydes til senere. Der stiles mod udskrivelse den []. Der har været et stigende behov for insulin, og vi kontakter endokrinologerne for råd herom.

eFigure 2. Overview of the convolutional neural network architecture.



eFigure 3. Overview of the recurrent neural network architecture.

eFigure 4. Overview of the hybrid architecture combining a convolutional and recurrent neural network.



eTable 1. Train, validation, and test distribution of the sentence level dataset.

	Training	Validation	Test	Total
Positive	4715	589	589	**5,893**
Negative	4715	589	589	**5,893**
Total	**9,430**	**1,178**	**1,178**	**11,786**

5.2 Improving and validating the bleeding identification model

In the previous paper *Deep Learning Detects and Visualizes Bleeding Events in Electronic Health Records* [18] we presented a bleeding detection algorithm based on a hybrid RNN-CNN architecture. To improve the results of that paper, we created a larger bleeding detection dataset and fine-tuned Clin-ELECTRA on the extended dataset.

In the follow-up paper *Doctors Identify Bleedings Better During Chart Review When Assisted by Artificial Intelligence* [19] we presented the extended dataset and the results of the fine-tuned Clin-ELECTRA model. Moreover, we evaluated MDs' accuracy in obtaining bleeding information with and without algorithm assistance in the narrative text of Danish EHRs. This was done to evaluate the potential benefits of using language technology in a clinical setting and to understand how it can assist MDs in their decision-making processes.

We found that the Clin-ELECTRA model performed with a sensitivity of 93.7% and a specificity of 98.1% on the test set. Moreover, we found that MDs identified more bleeding events during chart review when assisted by an algorithm and that they were positive towards using the algorithm for obtaining information regarding bleeding. This paper thereby indicated that pre-trained language models could be used as helpful assistive tools in clinical practice.

Due to copyrights, only the abstract is presented below.

Doctors identify hemorrhage better during chart review when assisted by artificial intelligence

Authors

- Martin Sundahl Laursen*,[1], MSc, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, msla@mmmi.sdu.dk
- Jannik Skyttegaard Pedersen*, MSc, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, jasp@mmmi.sdu.dk
- Rasmus Søgaard Hansen, M.D., Department of Clinical Biochemistry, Odense University Hospital, rasmus.sogaard.hansen@rsyd.dk
- Thiusius Rajeeth Savarimuthu, Professor, The Maersk Mc-Kinney Moller Institute, trs@mmmi.sdu.dk
- Rasmus Bank Lynggaard, MSc, Department of Clinical Biochemistry, Odense University Hospital, rasmus.bank.lynggaard@rsyd.dk
- Pernille Just Vinholt, Chief M.D., Department of Clinical Biochemistry, Odense University Hospital, pernille.vinholt@rsyd.dk

*Equal contribution

[1] Corresponding author, Campusvej 55, 5230 Odense M, Denmark, +4565504469

Abstract

Objectives: This study evaluated if medical doctors could identify more hemorrhage events during chart review in a clinical setting when assisted by an artificial intelligence (AI) model, and medical doctors' perception of using the AI model.

Methods: To develop the AI model, sentences from 900 electronic health records were labeled as positive or negative for hemorrhage and categorized into one of twelve anatomical locations. The AI model was evaluated on a test cohort consisting of 566 admissions. Using eye-tracking technology, we investigated medical doctors' reading workflow during manual chart review. Moreover, we performed a clinical use study where medical doctors read two admissions with and without AI assistance to evaluate performance when, and perception of using the AI model.

Results: The AI model had a sensitivity of 93.7% and a specificity of 98.1% on the test cohort. In the use studies, we found that medical doctors missed more than 33% of relevant sentences when doing chart review without AI assistance. Hemorrhage events described in paragraphs were more often overlooked compared to bullet-pointed hemorrhage mentions. With AI assisted chart review, medical doctors identified 48 and 49 percentage points more hemorrhage events than without assistance in two admissions, and they were generally positive towards using the AI model as a supporting tool.

Conclusions: Medical doctors identified more hemorrhage events with AI assisted chart review and they were generally positive towards using the AI model.

Keywords: Electronic Health Records, Hemorrhage, Artificial Intelligence, Decision Support Systems, Natural Language Processing

5.3 Development of a Danish clinical named entity recognition dataset

The two papers presented in this chapter deal with extraction of a specific medical condition. In the paper *Danish Clinical Named Entity Recognition and Relation Extraction* [20] we developed a NER dataset. In contrast to sentence classification tasks where a sample has a single label, NER samples have a label, known as a named entity, appended to all tokens in that sample, as seen in Figure 5.1. This is a challenging task because a single sample contains as many labels as there are words. Moreover, entities in the labeled corpus often follow a long-tail distribution meaning that a few entities account for a large part of the labels, while others appear only a few times [76].

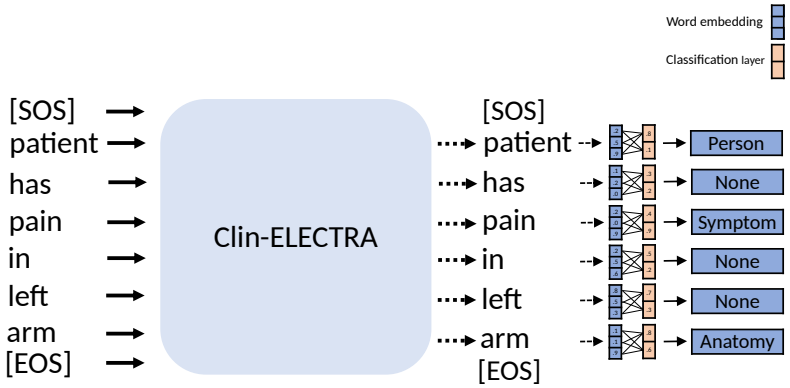


FIGURE 5.1: Illustration of named entity recognition where a prediction is performed for each word in the input sample. For example, 'patient' is predicted as the entity 'Person', and 'arm' is predicted as the entity 'anatomy'.

The developed NER dataset contains different types of entities such as diseases, symptoms, diagnostics, treatments, and test results. Moreover, each entity was described by an attribute, e.g., to specify if it is a current or prior disease. Finally, entities could also be linked by relations, e.g., to connect a diagnostic treatment with a result. This was done to make the dataset applicable to many different clinical scenarios.

The main challenge in developing a NER dataset is to create enough samples to create a well-performing algorithm, while at the same time ensuring a low rate of annotation errors. However, annotating a dataset for multiple entities and their connections is a challenging task. As written earlier, EHR text contains ambiguous expressions, variations in terminologies, and misspellings which make it difficult to identify and label entities correctly. Consequently, labeling each token in a sample with its corresponding named entity label is a labor-intensive and time-consuming task, especially when dealing with large datasets. Therefore, creation of a clinical NER dataset requires multiple medical professionals who are willing to spend hours on creating the dataset. To mitigate the risk of annotator errors, a crucial part of creating a NER

dataset is a thorough annotation process which requires creating detailed annotation guidelines, inter-annotator agreement checks, and motivated annotators [76]. Six annotators annotated the NER dataset consisting of 11,607 EHR paragraphs. Clin-ELECTRA was fine-tuned on the dataset and acquired a macro F1 of 60.05%, 44.85%, and 70.64% for clinical events, attributes, and relations, respectively. With further development, the developed model could for example be used to make a fast overview of a patient's disease and treatment history that the MDs could use to guide the treatment of the patient.

The paper is presented below. It is licensed under a CC BY 4.0 license which can be found at `https://creativecommons.org/licenses/by/4.0/`.

Danish Clinical Named Entity Recognition and Relation Extraction

Martin Sundahl Laursen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
msla@mmmi.sdu.dk

Jannik Skyttegaard Pedersen*
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
jasp@mmmi.sdu.dk

Rasmus Søgaard Hansen
Department of Clinical Biochemistry
Odense University Hospital

Thiusius Rajeeth Savarimuthu
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

Pernille Just Vinholt
Department of Clinical Biochemistry
Odense University Hospital

Abstract

Electronic health records contain important information regarding the patients' medical history but much of this information is stored in unstructured narrative text. This paper presents the first Danish clinical named entity recognition and relation extraction dataset for extraction of six types of clinical events, six types of attributes, and three types of relations. The dataset contains 11,607 paragraphs from Danish electronic health records containing 54,631 clinical events, 41,954 attributes, and 14,604 relations. We detail the methodology of developing the annotation scheme, and train a transformer-based architecture on the developed dataset with macro F1 performance of 60.05%, 44.85%, and 70.64% for clinical events, attributes, and relations, respectively.

1 Introduction

Electronic health records (EHR) contain important information regarding the patients' medical history including diagnoses, medications, treatment plans, allergies, and test results. However, much of this information is stored in unstructured narrative text. While this information could be used to guide diagnostic decision making and treatment plans, the unstructured format makes it infeasible to fully exploit in clinical practice and research.

Natural language processing (NLP) algorithms could be used to transform the unstructured narrative text of the EHR into structured information and give medical doctors (MD) a fast overview of even a medical history spanning multiple years. NLP models' ability to process and extract information from written text keeps improving with benchmark-breaking models being published on a regular basis. For example, transformer-based models such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), and ELECTRA (Clark et al., 2020) have recently shown promising results for many NLP tasks, e.g. named entity recognition and relation extraction (NER). In NER, models are trained to tag words with predefined entities and find the relations between them. In clinical NER, entities such as diseases, treatments, drugs, and tests have been extracted automatically from EHRs. However, many of the developed datasets are only in English and for specific clinical specialities or note types (Uzuner et al., 2007, 2010; Bethard et al., 2016).

This paper describes the methodology for developing the first Danish clinical NER dataset. The dataset consists of text paragraphs from Danish EHRs spanning multiple departments and note types.

First, the paper describes the clinical dataset, the strategy for choosing entities tailored to extract important information from EHRs, and the annotation scheme. Next, we train a transformer-based architecture on the developed NER dataset.

2 Related works

The annotation schemes and extracted clinical entities and relations vary. Agrawal et al. (2022) extracted medications, their status (active, discontinued, neither), and attributes. The i2b2 2009 challenge (Uzuner et al., 2010) and n2c2 2018 Track 2 (Henry et al., 2020) only extracted medications

*Equal contribution

Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 655–666

and their attributes. Examples of attributes are name, dosage, mode of administration, frequency, duration, reason, strength, form, and adverse drug effects.

SemEval-2016 Task 12 (Bethard et al., 2016) extracted time entities; event entities and their contextual modality, degree, polarity, and type; and temporal relations between time and event entities (before, overlap, before—overlap, after).

SemEval-2015 Task 14 (Elhadad et al., 2015) and CLEF eHealth 2013 Task 1 (Pradhan et al., 2015) extracted disorder mentions and mapped them to their UMLS/SNOMED concept unique identifier. The former also classified attributes such as the disorder's subject, course, body location, and severity, and whether it was negated, uncertain, conditional, or generic.

The i2b2 2010 challenge (Uzuner et al., 2011) extracted entities (medical problems, treatments, tests), assertions (present, absent, possible, conditional, hypothetical future, and associated with someone other than the patient), and relations between medical problem entities and each of medical problem, treatment, and test entities.

The i2b2 2012 challenge (Sun et al., 2013b) extracted clinically relevant events. Their type was classified as concept (problem, test, treatment), clinical department, evidentials indicating source of information, or occurrences (events that happen to the patient). Polarity was classified as positive or negated, and modality as happens, proposed, conditional, or possible. Temporal expressions were extracted with their type (date, time, duration, frequency), value, and modifier indicating whether the temporal expression was exact or not. Temporal relations indicating the type of connections between events and temporal expressions were also extracted.

3 Methods

This section describes the data, annotation scheme, and model used for Danish clinical NER.

3.1 Data

We extracted 11,607 paragraphs with a length between 11 and 75 words from EHRs from Odense University Hospital in Denmark. Paragraphs were sampled randomly from different EHR note types across every department of the hospital to ensure the data distribution would resemble that of EHRs: 46% were from clinical contacts, 13% primary journals, 10% care data, 3% epicrises, 3% ambulatory care contacts, 2% surgical notes, 2% emergency room journals, and 20% were from 55 different minor EHR note types. Paragraphs were lowercased and anonymised by two of the authors.

Clinical event	Description
Disease	A disorder of structure or function, especially one that has a known cause and a distinctive group of symptoms, signs, or anatomical changes. Examples include cancer, influenza, and narcolepsy.
Symptom	A symptom is a physical or mental feature which is regarded as indicating a condition of disease, particularly such a feature that is apparent to the patient. We include abnormal findings, which the MD makes when examining the patient objectively, as these are sometimes coinciding with symptoms—e.g. bruises. Examples include headache, stomach ache, and pain.
Diagnostic	Any tool or method concerned with the diagnosis of illnesses or other problems. Includes measurements and tests. Examples include CT scans, blood samples, and temperatures.
Treatment	A treatment is any medical care given to a patient for an illness or injury. Examples include medication, plaster, and rehabilitation.
Anatomy	Any part of human anatomy. Includes body fluids and excrements. Examples include arms, organs, and blood.
Result	All results of diagnostics that do not carry any meaning without being coupled to the diagnostic. Examples include numbers that indicate length, temperature, or volumes. Diseases or symptoms found by diagnostics are annotated as such, e.g. a tumour found by a CT scan.

Table 1: Description of clinical events. Descriptions were inspired by the Oxford English Dictionary.

3.2 Annotation

3.2.1 Annotation scheme

Two MDs with expert clinical domain knowledge developed the annotation scheme through an iterative process of making annotation rules and testing them.

Annotation rules were made to extract clinically relevant information from the medical history. Focus was for the rules to be as complete as possible to capture all important information about the medical history while still being simple to use for the annotators.

We extracted three types of information: clinical events, the attributes of the clinical events, and relations between the clinical events.

Clinical events were: diseases; symptoms, including abnormal findings; diagnostics; treatments; anatomies including body fluids and excrements; and results. Symptoms and abnormal findings were joined in one as they sometimes coincided. Normal findings were not included as there were so many that they would cloud the visualisation of the history. Table 1 shows all clinical events and their descriptions as defined by the medical experts.

Clinical events were further described by their attributes. Attributes were: prior; current; fu-

Attributes	Description
Prior	Entities that occurred in prior admissions or in the distant past. Includes treatments that are being stopped at that point in time.
Current	Entities that occur in the present. Includes prescribed medicine.
Future	Entities that occur or might occur in the future—e.g. the risk of skin cancer, or ordering diagnostics for a later day.
Doubt	Any entity that is not confirmed. Includes any treatments that might need to be started in the future.
Negation	Entities such as diseases or symptoms that are mentioned as not being present.
Non-patient	Entities that are not related to the patient in question. One example is the disease history of the patient's relatives.

Table 2: Description of attributes.

ture; doubt; negation; and non-patient. All clinical events could take one of the six attributes except anatomies and results. Anatomies did not take any attributes while results could only take a prior or current attribute. Table 2 shows all attributes and their descriptions.

Clinical events could connect to each other in limited ways through one-way relations. Diseases, diagnostics, and symptoms could connect to anatomies through a "has location" relation. Diseases, symptoms, and anatomies could connect to treatments through a "is treated with" relation. Diagnostics could connect to results through a "has result" relation.

Figure 1 shows an overview of the clinical events, attributes, and relations. Appendix A shows the full annotation guidelines with further details and explanations to the annotators.

3.2.2 Annotation process

Six annotators were recruited for the task. Five were Master of Science in Medicine students and one was a MD.

Figure 2 shows the process of annotator training. It included reading the annotation guide and an iterative process of annotating a learning set of 55 paragraphs (not included in dataset) followed by error analysis until a final test was made on a set of 98 gold paragraphs annotated by an expert MD. Paragraphs were annotated using the CLAMP software (Soysal et al., 2017). We report the micro F1 of each annotator on the gold set.

Figure 3 shows an example of an annotated paragraph.

3.3 Entity and relation extraction model

This section describes the architecture of the Princeton University Relation Extraction system (PURE) (Zhong and Chen, 2021) which we used and adapted for Danish clinical NER. It further describes the dataset used and the training of the models.

3.3.1 Model architecture

PURE—the 2021 state-of-the-art on entity and relation extraction—is a NER deep learning model based on a transformer structure. The model has a separate entity and relation extraction part.

For entity extraction, the model takes as input all possible text spans up to a maximum length. A transformer extracts contextual word embeddings for the start and end token of each span. They are concatenated with a learned span width embedding and classified by a feedforward network.

When extracting relations, for each candidate pair of entities, the text is passed through a transformer with inserted entity start and end marker tokens for the subject and object entity, also indicating the type. The concatenation of the start marker token for the candidate subject and object entity is classified by a feedforward neural network.

We used PURE's entity extraction approach for clinical events and the relation extraction approach for relations between clinical events.

We used our own approach adapted from the PURE relation extraction approach for attributes. We inserted clinical event start and end marker tokens, passed all tokens through a transformer, concatenated the start and end marker tokens, and classified the attribute using a feedforward network. The marker tokens were used for classification instead of the word(s) forming the clinical event to guide the model to look more at the context rather than the specific word—the context being the important factor in attribute classification. Additionally, enriching the input with the type of the clinical event could guide the model if attributes were described differently for different clinical events.

Figure 4 shows the three types of extraction tasks.

3.3.2 Datasets

Table 3 shows the number of clinical events, attributes, and relations by type in the train, validation, and test set. The dataset had a total of 11,607 paragraphs, each containing a varying number of clinical events, attributes, and relations. On average, each paragraph contained 4.7 clinical events, 3.6 attributes, and 1.3 relations. We split the paragraphs in train, validation, and test sets for an approximate 80%–10%–10% ratio between each type of clinical event, attribute, and relation. The sets were unbalanced on type of entity or relation—e.g. for the attributes training

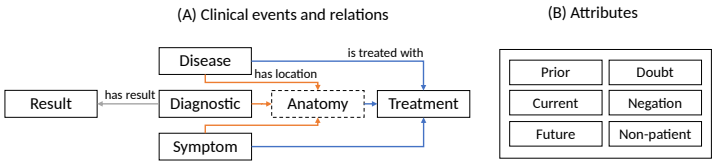


Figure 1: (A) Clinical events and relations between them. Symptoms include abnormal findings. Anatomies include body fluids and excrements. Diagnostics include measurements and tests. Blue: "is treated with". Orange: "has location". Grey: "has result". (B) Attributes. Anatomy (dashed lines) takes no attributes. Other clinical events must take one attribute. Results only take prior or current attributes.

	Train (% of row total)	Validation (% of row total)	Test (%of row total)	Total (% of column total)
Paragraphs	9,687 (83%)	960 (8%)	960 (8%)	**11,607 (100%)**
		Clinical events		
Diseases	2,033 (78%)	295 (11%)	272 (10%)	**2,600 (5%)**
Symptoms	11,937 (80%)	1,455 (10%)	1,571 (10%)	**14,963 (27%)**
Diagnostics	8,921 (80%)	1,095 (10%)	1,194 (11%)	**11,210 (21%)**
Treatments	6,918 (79%)	911 (10%)	882 (10%)	**8,711 (16%)**
Anatomies	10,172 (80%)	1,227 (10%)	1,278 (10%)	**12,677 (23%)**
Results	3,522 (79%)	473 (11%)	475 (11%)	**4,470 (8%)**
TOTAL	**43,503 (80%)**	**5,456 (10%)**	**5,672 (10%)**	**54,631 (100%)**
		Attributes		
Prior	2,028 (80%)	237 (9%)	283 (11%)	**2,548 (6%)**
Current	23,217 (79%)	3,021 (10%)	3,109 (11%)	**29,347 (70%)**
Future	1,237 (79%)	161 (10%)	160 (10%)	**1,558 (4%)**
Doubt	2,479 (82%)	263 (9%)	289 (10%)	**3,031 (7%)**
Negation	3,890 (80%)	496 (10%)	500 (10%)	**4,886 (12%)**
Non-patient	480 (82%)	51 (9%)	53 (9%)	**584 (1%)**
TOTAL	**33,331 (79%)**	**4,229 (10%)**	**4,394 (10%)**	**41,954 (100%)**
		Relations		
is treated with	1,485 (80%)	175 (9%)	197 (11%)	**1,857 (13%)**
has location	6,501 (80%)	779 (10%)	823 (10%)	**8,103 (55%)**
has result	3,652 (79%)	499 (11%)	493 (11%)	**4,644 (32%)**
TOTAL	**11,638 (80%)**	**1,453 (10%)**	**1,513 (10%)**	**14,604 (100%)**

Table 3: Composition of the train, validation and test sets by type of clinical event, attribute, and relation.

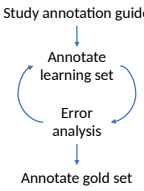


Figure 2: Annotator training process. Figure inspired by Sun et al. (2013a).

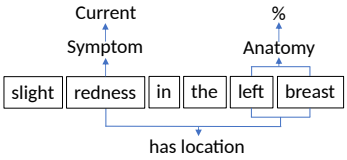


Figure 3: Example of annotated paragraph. % signifies that no attribute could be assigned to the clinical event per the annotation scheme.

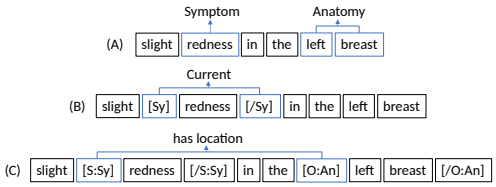


Figure 4: (A) Classification of clinical events from start and end tokens of span. Span width embedding not depicted. (B) Classification of attribute using clinical event marker tokens. (C) Classification of relation using subject/object and clinical event marker tokens. Figure inspired by Zhong and Chen (2021).

Evaluation metric	Loss	Micro			Macro		
		R%	P%	F1%	R%	P%	F1%
Micro F1	Unweighted	79.14	79.14	79.14	38.34	40.51	38.56
	Weighted	61.81	61.81	61.81	45.35	33.20	34.23
Macro F1	Unweighted	77.30	77.30	77.30	41.88	41.90	41.48
	Weighted	60.13	60.13	60.13	51.37	41.87	43.85

Table 4: Validation set micro and macro recall, precision, and F1 score on the attribute extraction task when selecting the best iteration of the model based on micro and macro F1 score with unweighted and weighted loss. 2 hidden layers of size 75 was used for the test. R: Recall. P: Precision.

set, there were 23,217 current and only 480 non-patient attributes. All datasets were in the json format used by PURE (see Zhong and Chen (2021)).

3.3.3 Training

When training the clinical event extraction model, we used a Danish Clinical ELECTRA pretrained on the narrative text from 299,718 EHRs from Odense University Hospital as the transformer base (Pedersen et al., 2022). The model had ∼13M parameters and consisted of 12 transformer layers with 4 attention heads. We used a dropout of 0.1 after the last ELECTRA hidden layer output. We tested classification heads with two hidden layers of varying size, each followed by a dropout of 0.2 and a ReLU activation function. We used a maximum span of 8 and a train batch size of 32. We trained for 100 epochs using the AdamW optimizer with learning rate 1e-5 for the transformer layers and 1e-4 for the classification head, and a warm-up proportion of 0.1.

When training each of the models for extracting attributes and relations, we used the same transformer base with a normalisation layer and a dropout of 0.1 after the concatenation of tokens. We tested classification heads with two hidden layers of varying size, each followed by a dropout of 0.2 and a ReLU activation function. We further tested a classification head only consisting of a single classification layer. We used a train batch size of 32 and a maximum sequence length of 128. We trained for 20 epochs using the AdamW optimizer with learning rate 2e-5 and a warm-up proportion of 0.1.

We modified the training method of PURE to guide the models towards equal performance on all classes by using a weighted loss function to counteract the unbalanced dataset and chosing the best model for each of the clinical event, attribute,

and relation extraction tasks as the model iteration with the best macro F1 on the validation set, rather than the micro F1 standard of PURE. Table 4 shows a test of the performance on the attribute extraction task when selecting the best iteration of the model based on micro and macro F1 score with unweighted and weighted loss. Using the macro F1 score with weighted loss gave the best performance across all classes. Appendix B shows the confusion matrices for each combination.

Class weights were calculated for the training of each model using the default formula in Scikit-learn (Pedregosa et al., 2011):

$$w_x = \frac{n_{samples}}{n_{classes} \cdot n_x} \quad (1)$$

where x is the class, $n_{samples}$ is the number of total samples, and $n_{classes}$ is the number of classes. The negative class, i.e. samples not to be given any label by the model, was given a weight of 1.

The negative class was excluded when calculating the F1. We only trained the attribute and relation models to make classifications that were allowed for the connected clinical events according to the annotation scheme. Appendix C shows the results of the hyperparameter search. We report the micro and macro recall, precision, and F1 for the best models on the test set.

4 Results

This section presents the agreement of the annotators on the gold set and the results of the Danish clinical NER models.

4.1 Annotation

Table 5 shows the annotators' micro F1 performance on the gold set. For clinical events, it ranged 83.71%–91.24% (average 85.62%) for overlapping matches, and 74.12%–85.15% (average 77.67%) for exact matches. For attributes, it ranged 79.21%–86.19% (average 81.71%) and for relations 71.28%–90.06% (average 77.79%).

4.2 Entity and relation extraction model

The models that had the best validation performance in the hyperparameter search were:

- A clinical event extraction model with two hidden layers of size 450 in the classification head.
- An attribute extraction model with a single classification layer.
- A relation extraction model with two hidden layers of size 150 in the classification head.

Annotator	A	B	C	D	E	F
	Overlap match, micro F1%					
Clinical event	91.24	84.22	84.41	85.71	84.43	83.71
Attribute	86.19	83.06	79.21	81.29	79.75	80.75
Relation	90.06	76.97	75.60	77.01	71.28	75.84
	Exact match, micro F1%					
Clinical event	85.15	76.08	76.29	78.69	74.12	75.71

Table 5: The anonymised annotators' performance on the gold set. Exact match: a match is defined as the exact tokens annotated in the gold set with the same label. Overlap match: a match is defined as minimum one token overlapping with the gold set annotation of the same label. Only an overlap match F1 is calculated for attributes and relations as evaluating an exact match would propagate the potential error in the span of the clinical event to which the attribute or relation is connected.

	Micro			Macro		
	R%	**P%**	**F1%**	**R%**	**P%**	**F1%**
	Overlap match					
Clinical events	66.29	77.31	71.38	64.88	72.60	68.20
	Exact match					
Clinical events	60.97	65.64	63.22	59.84	61.30	60.05
Attributes	66.04	66.04	66.04	51.60	42.64	44.85
Relations	75.88	72.66	74.23	74.74	67.85	70.64

Table 6: Performance of the best clinical event, attribute, and relation extraction models on the test set. Attributes and relations are only reported with an exact match as the models do not consider the span of the clinical event from which the attribute or relation is classified. R: Recall. P: Precision.

Table 6 shows the performance of the best models on the test set. Clinical events were extracted with exact micro F1 63.22% and macro F1 60.05%, attributes with micro F1 66.04% and macro F1 44.85%, and relations with micro F1 74.23% and macro F1 70.64%. The negative class was excluded when calculating the recall, precision, and F1 scores.

Figure 5 shows the confusion matrices of performance on clinical events, attributes, and relations. The confusion matrices include the clinical events and relations that were not extracted and falsely extracted by the model ('O').

The model for clinical event extraction performed best on anatomies (69%) and worst on results (53%). 1,568 spans were falsely extracted

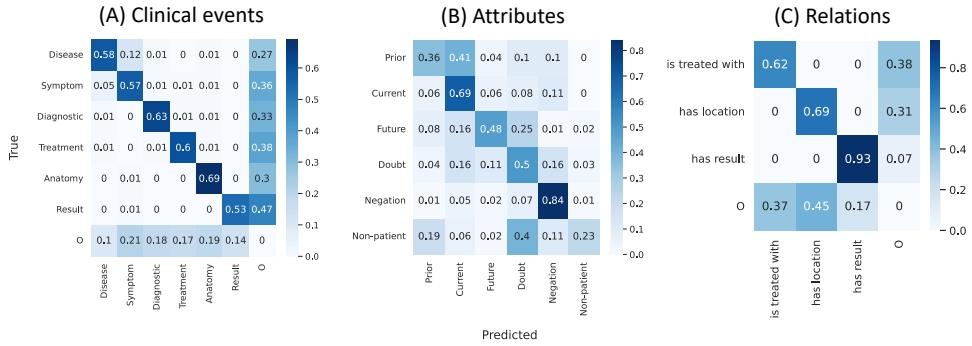


Figure 5: Confusion matrices of performance on (A) clinical events, (B) attributes, and (C) relations. 'O' counts the clinical events and relations that were not extracted and falsely extracted by the model.

as a clinical event with symptoms being the most frequent (21%). The model for attribute extraction performed best on negations (84%) and worst on non-patient (23%). The model for relation extraction performed best on "has result" (93%) and worst on "is treated with" (62%). 432 false relations were extracted of which "has location" was the most frequent misclassification (45%).

5 Discussion and limitations

This paper presented a methodology for developing a dataset for Danish clinical NER. It presented an annotation scheme for annotation of all clinical events, their attributes, and relations that are relevant for the medical history. The dataset included text paragraphs from Danish EHRs spanning multiple departments and note types.

We trained and adapted PURE NER deep learning models to extract clinical events (overlap match macro F1 68.20%; exact match macro F1 60.05%), attributes of clinical events (macro F1 44.85%), and relations between clinical events (macro F1 70.64%). The results are promising for Danish clinical NER but need improvement. A discussion of possible improvements to the methodology, limitations, and future work is provided below.

The clinical event extraction model had similar performance on all classes with accuracies between 53% (results) and 69% (anatomies). There was little contamination between classes as most errors were caused by failure to extract or false extraction of a clinical event. There was some contamination between symptoms and diseases with 12% of diseases being classified as symptoms and 5% of symptoms being classified as diseases. This supports claims by annotators that diseases and symptoms in some cases are difficult to differentiate and that extra attention must be given to differentiate these in the annotation guidelines.

The attribute extraction model had large differences in performance with accuracies between 23% (non-patient) and 84% (negation). There were more misclassifications of the non-patient attribute as doubt (40%) than correct classifications. The future and doubt attributes had significant contamination between them with 25% and 11% misclassifications as the other class, respectively. The many misclassifications between non-patient and doubt attributes, and especially future and doubt attributes, could indicate that the model would improve if the non-patient, doubt, and future attributes were merged to a single class of uncertain attributes. This would most likely not harm the usefulness of the model to MDs significantly.

The fact that more prior attributes were misclassified as current (41%) than correct classifications (36%) likewise indicates that these two attributes could be merged into a single class of clinical events that occurred. This would, however, decrease the usefulness of the model as it is important for MDs reviewing the medical history to know if a clinical event is prior or current.

The relation model extracted 93% of the "has result" relations, and 62% and 69% of the "is treated with" and "has location" relations, respectively. The differences are likely caused by the fact that the "has result" relation only connects diagnostics to results while the two other relations have three different one-way relationships.

In this paper, we only explored one type of NER model and tested a limited set of architectures and hyperparameters. Future work could include testing other architectures and enriching the model input with more information, e.g. the output of a text parser, which could help differentiate attributes dealing with the time-aspect. The six annotators had an average micro F1 (overlap match) of 85.62%, 81.71%, and 77.79% for clinical events, attributes, and relations, respectively. Merging certain attributes and more emphasis on differences between symptoms and diseases could increase these scores.

The Danish clinical NER dataset is not made publicly available due to it containing sensitive information. We advise interested researchers to contact us for sharing possibilities.

6 Conclusions

This paper presented methodology and annotation scheme for developing the first Danish clinical NER dataset. The corpus consists of 11,607 paragraphs annotated for six entity types, six attributes, and three relations. The corpus was used to fine-tune language models which showed promising results for classifying the entities, attributes, and relations of the dataset.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 1052–1062.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. https://openreview.net/forum?id=r1xMH1BtvB Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. Semeval-2015 task 14: Analysis of clinical text. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. https://doi.org/10.1093/jamia/ocx132 CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. https://doi.org/https://doi.org/10.1016/j.jbi.2013.07.004 Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46:S5–S12. Supplement: 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61.

Appendices

A Annotation guidelines

A.1 Clinical events

A.1.1 Disease

Contains all diseases including diseases that could be considered a result of a Diagnostic.

A.1.2 Symptom

Includes all symptoms and abnormal findings. Findings that are not abnormal should not be annotated. However, a negation of an abnormal finding should be annotated because the abnormal finding is mentioned even though it is not present. For example, "fracture" should be annotated in the sentence "there is no sign of fracture."

If there is a negation of a non-abnormal finding, it should be annotated in the entity. For example, "cannot hear" is annotated in the sentence "patient cannot hear anything."

In the sentence "no symptoms," the word "symptoms" should not be annotated as a symptom, as it does not contain any information.

In case a symptom or abnormal finding is found by a Diagnostic, there may be a coincidence with the Result entity. Here, it is annotated as Symptom if the entity can provide sufficient meaning alone. For example, "cyst" or "tumour."

If the Symptom cannot stand alone and one needs to know which Diagnostic was carried out in order to understand the result, the entity should instead be annotated as Result and have a "has result" relationship from the Diagnostic entity. For example, this applies to "Temp: 24 C" and "Stix: 3+". "Temp" and "Stix" are annotated as Diagnostic with "is treated with" relationship to Result "24 C" and "3+."

A.1.3 Result

Includes all results of Diagnostic, e.g. values and blood test results.

A Result cannot stand on its own. A relation from the Diagnostic is needed for it to make sense. These can be entities like "stable", "positive", "negative", "24 C" or "3+".

Typically, this entity will appear in sentence structures with a colon: "Diagnostic: Result". Note that the two entities are mentioned very close to each other in the text—in this case only with a colon in between. An example could be "Temp: 24 C" or "Stix: 3+". "Temp" and "Stix" are annotated as Diagnostics with a "has result" relation to Result "24 C" and "3+".

Entities that can instead be annotated as Symptom will typically be mentioned further away or completely lack a Diagnostic as a Symptom can stand alone and make sense.

See also the description for Symptom.

A.1.4 Diagnostic

Includes all diagnostics, measurements, and tests. This can include CT scans, blood tests, MR scans, and recordings of a newborn's length, temperature, etc.

Note that "blood sample results" and "radiology description" are not a Diagnostic and should not be annotated.

If KAD is mentioned along with a volume, e.g. "KAD emptied of 200 mL," it is marked as Diagnostic - Result. If there is no volume specified, KAD is annotated as Treatment.

A.1.5 Treatment

Includes all forms of treatment including medication.

To annotate entities as concisely as possible, for example in the sentence "good effect of 2.5 mg morphine IV," only "morphine" should be annotated as Treatment.

In the sentence "treated for xxx," the word "treatment" should not be annotated as Treatment as it does not contain any information.

If KAD is mentioned without a volume indication, it should be annotated as Treatment. If KAD is mentioned with a volume, for example "KAD emptied for 200 mL," it should be annotated as "Diagnostic - Result."

A.1.6 Anatomy

Includes all mentions of anatomies and things from the body (blood, feces, urine, sweat, etc.).

Typically used to indicate the location of a Disease or Symptom, a Diagnostic, or a Treatment. Examples: "brain", "left foot" or "duodenum".

When Anatomy is described by an adjacent word, for example "left", this should be included in the entity.

Remember to annotate the Anatomy entities that should not be linked to other entities.

A.2 Attributes

A.2.1 Current

The entity is either present, carried out, or current. If medication is prescribed to the patient, this should also be marked as "Treatment - Current", as it can be assumed that the treatment will start and it may be the last time it is mentioned in the journal. On the other hand, "Scheduling a CT for Tuesday." should be marked as "Future" as it will be described in a future medical note, for example with the result.

A.2.2 Negation

The entity is not present. For example, if it is mentioned that the patient does not have a fracture, the fracture should be marked as Symptom - Negation. Note that the word "not" should not be part of the marked entity. However, if there is a negation of a normal finding, it should be annotated as such. For example, "cannot hear" in the sentence "patient cannot hear anything" is annotated as Symptom - Present.

A.2.3 Prior

If the entity refers to a previous case, i.e., a previous hospitalisation or if it happened a long time ago. For example, it should be annotated as a prior Treatment when a cast or drain is removed, as the treatment is finished. However, if a CT scan from the previous day is mentioned, it should be annotated as Current.

A.2.4 Future

Everything that takes place in the future. For example, cancer is annotated as Disease - Future if it is mentioned that "there is a risk of cancer if you use tanning beds too often."

It is marked as Diagnostic - Future if an MRI scan is planned for the next day. However, if it is written "the treatment with xxx starts" or "rp. xxx" it should be marked as Treatment - Current as it is assumed that the treatment will certainly happen.

Also includes references to possible future treatments.

A.2.5 Doubt

If the patient might have a disease that has not yet been confirmed.

If a Treatment should be given provided that certain things change.

The difference between Doubt and Future is that Future is more certain - it is going to happen - while Doubt is more uncertain or conditional.

A.2.6 Non-patient

If an entity does not have a direct connection to the patient. This can occur when a general letter is sent out regarding cancer screening. Cancer should then be annotated as Disease - Non-patient. If it is mentioned that the patient's mother had a certain disease, it should also be annotated in this way.

A.3 Relations

When entities are annotated, the relationships between entities can be annotated. This is done by pulling the "From entity" over to the "To entity". The direction of the relationship is important. Therefore, pay attention to the name of the relationship and read it out loud if necessary, "Entity - Relation - Entity" and listen to see if it makes sense or if the arrow needs to be reversed. CLAMP will show which relationships can be annotated for the pair being drawn between.

has location

From entities: Disease, Symptom, Diagnostic.
To entities: Anatomy.

has result

From entities: Diagnostic.
To entities: Result.

is treated with

From entities: Disease, Symptom, Anatomy.

To entities: Treatment.

The "is treated with" relation links the entities Disease, Symptom, and Anatomy to a Treatment. In some cases, sentences describing a required treatment could be linked to both an Anatomy and Treatment entity. In this case, the Treatment should be linked to the Symptom instead of the Anatomy. You should only link the Anatomy to the Treatment using the "is treated with" relation if the Treatment cannot be linked to anything else. Example: "Left knee skin scraping is treated with plaster." Annotation: skin scraping - "Treated with" - plaster.

A.4 General notes

It is important not to annotate periods, commas, etc. unless they are part of an abbreviation. For example, in "Patient has cancer," only "cancer" and not "cancer." should be marked. If you double-click a word, CLAMP will only mark the word and not any punctuation next to the word. This can make it a bit troublesome to include periods in abbreviations.

Entities should be annotated as concisely as possible without losing meaning. This means that in the sentence "there are signs of cancer," only "cancer" and not "signs of cancer" should be marked as an entity. If an entity has some describing words next to it, the following rule can be used to decide how much should be annotated. In the sentence "pain in the front of the arm," only "arm" is marked as Anatomy since "front" and "arm" are connected through the word "of." In the sentence "pain in the left arm," "left arm" is marked as Anatomy since there are no words between "left" and "arm". In sentences describing a prescription of medication, only the name is marked as Treatment, and not, for example, the quantity indication or the number of days.

Entities may not overlap with each other.

B Selection of loss and evaluation metric

Figure 6 shows the confusion matrices for the attribute extraction task when selecting the best iteration of the model based on micro and macro F1 score with unweighted and weighted loss.

Using the micro F1 to select the best iteration of the model resulted in some classes being practically excluded during classification. Using the macro F1 to select the best model iteration and training with a weighted loss gave the most equal performance on all classes.

	Classification head hidden layers	Validation Exact F1 %
Clinical event	2x 75	58.49
	2x 150	59.82
	2x 300	60.68
	2x 450	61.34
	2x 600	60.91
Attribute	None	48.01
	2x 50	43.20
	2x 75	43.85
	2x 150	44.10
	2x 300	44.32
Relation	None	66.15
	2x 75	68.39
	2x 150	68.85
	2x 300	67.39

Table 7: Results of the hyperparameter search.

C Hyperparameter search

Table 7 shows the results of the hyperparameter search.

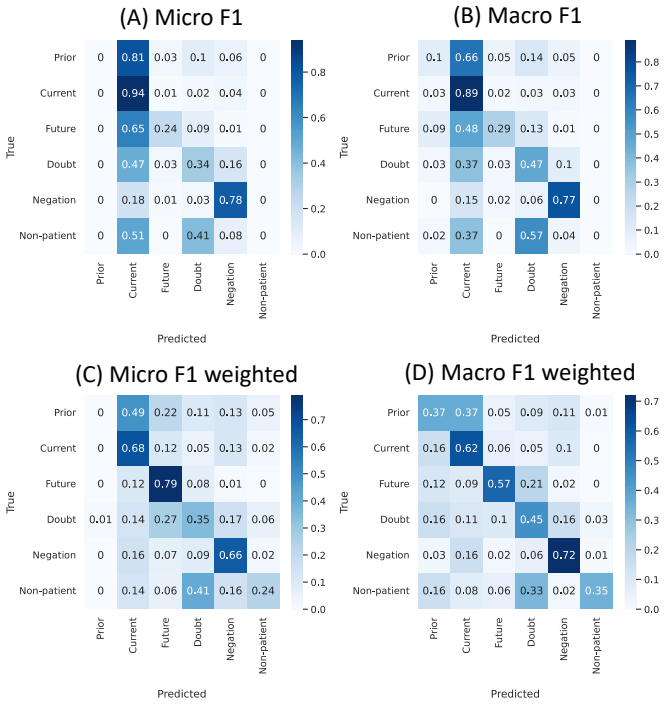


Figure 6: Confusion matrices showing the attribute extraction validation performance of the models chosen based on (A) micro F1, (B) macro F1, (C) micro F1 trained with weighted loss, and (D) macro F1 trained with weighted loss.

CHAPTER 6

Conclusion

The work during my studies has been concerned with developing Danish clinical language technology resources by 1) collecting large EHR datasets to pre-train language models, 2) developing pre-trained language models, 3) developing datasets that can be used to fine-tune language models and extract different diseases and symptoms in the narrative text of Danish EHRs, and 4) evaluate and compare the language models with regards to accuracy and possible biases in clinical classification tasks.

This chapter revisits the research questions of the dissertations by reviewing how each chapter contributed to answer these.

- **How do Danish clinical pre-trained language models perform compared to Danish open-source pre-trained language models and other deep learning-based architectures?**

Chapter 3 presented our work on developing two pre-trained language models, Clin-ELECTRA and MeDa-BERT, that can be used for Danish clinical NLP tasks.

Clin-ELECTRA was pre-trained on a large clinical corpus of approximately 300,000 Danish EHRs. The performance of the Clin-ELECTRA model was compared with a general domain BERT model and an LSTM model with clinical word embeddings. We found that Clin-ELECTRA performed better than the two other models, even though the BERT model was ten times larger in terms of parameters.

MeDa-BERT was pre-trained on Danish medical text from the internet and medical books. The MeDa-BERT model performed better on four clinical classification tasks compared to a Danish BERT model trained on the general domain. The collected corpus was aslo used to develop medical word embeddings. When used as input to an LSTM model, the medical word embeddings performed better than general domain word embeddings. The medical word embeddings and the MeDa-BERT model were made publicly available to strengthen development and research in Danish NLP.

These results suggest that Danish clinical language models are valuable for extracting information from Danish EHRs as their performance surpass other pre-trained language models and deep learning-based architectures.

- **How can bias in clinical language models arise from the representation of patient subgroups in the datasets used for model development?**

Chapter 4 presented the paper *Investigating anatomical bias in clinical classification algorithms* [17]. This paper highlights the crucial role of dataset curation when developing language models. We found that the performance of language models for classifying bleeding and thrombosis varied significantly for specific patient populations if the training samples were not evenly distributed across anatomical locations. The paper stressed the importance of carefully constructing datasets that are used to train language models in order to avoid anatomical bias. This is essential to avoid implementing clinical algorithms that might discriminate against specific patient populations.

- **How can clinical language models contribute to extracting clinically important information from Danish EHRs?**

Chapter 5 presented three papers that focus on extracting information from the narrative text of Danish EHRs.

The paper *Deep learning detects and visualizes bleeding events in electronic health records* [18] presented our initial work on extracting bleeding episodes from Danish EHRs. In this study, we created a large annotated dataset consisting of paragraphs from EHRs that did or did not describe bleeding events. This dataset was used to train a hybrid RNN-CNN model with clinical word embeddings to classify paragraphs as either positive or negative for bleeding. The model performed with an accuracy of 90% on the test set.

In our follow-up paper, *Doctors identify bleedings better during chart review when assisted by artificial intelligence* [19], we used Clin-ELECTRA to improve our accuracy for bleeding classification, and we presented a user study that investigated MDs' performance when using the algorithm in a clinical scenario and the MDs perception of using it. MDs had a positive sentiment toward using the algorithm in clinical practice, and they registered more bleeding episodes when using the bleeding classification algorithm. We found that MDs failed to register up to 46% of bleeding paragraphs when reviewing an admission without algorithm assistance. This indicates MDs may not obtain the information they need for the assessment of bleeding risks, and that algorithm assistance could provide important information to MDs.

In the paper *Danish Clinical Named Entity Recognition and Relation Extraction* [20], we presented a Danish clinical NER dataset with six types of clinical events, six types of attributes, and three types of relations. The Clin-ELECTRA model was fine-tuned on the dataset and achieved a macro F1 of 60.1%, 44.9%, and 70.6% for clinical events, attributes, and relations, respectively. Contrary to the bleeding dataset, the NER dataset can be used to develop language models that can provide valuable information in many different clinical scenarios, e.g., to get an overview

of a patient's disease history.

In summary, this dissertation has developed resources for Danish clinical NLP, contributed with research for mitigating bias in clinical language models, and shown promising results for extracting clinically important information from the narrative text of Danish EHRs. The findings of the dissertation suggest that Danish clinical pre-trained language models can play a significant role in shaping the future EHR system. Today, EHRs are designed to facilitate the retrieval of clinical information by structuring the information as much as possible. However, a large part of the EHR consists of unstructured narrative text. Consequently, when the number of healthcare contacts increases for each patient, there is a risk of information overload causing crucial clinical information to be lost in the narrative text. Language technology could help mitigate the problem of information overload by facilitating effective search functionalities within EHR systems. This would enable healthcare professionals to efficiently navigate through the EHR, reduce the time-expensive task of manually reviewing the EHR, and free up time for the MD to direct patient interaction. Additionally, it could advance clinical research possibilities, as access to relevant clinical data becomes more readily available. To fulfill these positive outcomes, it is crucial to develop safe and unbiased language models. Language models could have a direct effect on all patients in the institution it is utilized, and therefore their outcome predictions must be thoroughly evaluated on the patient groups they are meant to assist.

6.0.1 Future work

Throughout my PhD studies, language technology has advanced at a rapid pace with new innovations, possibilities, and challenges emerging every month. Often, while working on implementing and acquiring knowledge about a new technology, a new and better technology has already been developed. Since the models and datasets presented in this dissertation were developed, new research has been introduced. This sub-section presents the potential directions and areas of research that could build upon the findings of the dissertation and contribute to further advancement in Danish clinical NLP.

Generalization across Danish hospitals

The models developed for this dissertation have only been trained and tested on data from OUH. In the future, we will evaluate the models on datasets with EHR text from other hospitals in Denmark. This is necessary since other hospitals could use terminologies, abbreviations, and linguistic patterns different from that of OUH. By testing the models on text from other hospitals, we can assess their ability to adapt and perform consistently across these linguistic differences. This will provide insights into the models' robustness and generalization abilities.

Development of new pre-trained clinical language models

When introduced, the ELECTRA pre-training approach was the most efficient method for pre-training language models. After its introduction, other innovations quickly followed. For example, the DeBERTa model [77] made small improvements to the original transformer architecture by using a new attention mechanism where each token is represented using two vectors that encode its content and position, and a new way to capture the absolute position of tokens. In a follow-up paper, the authors presented DeBERTaV3 [78] that combined the DeBERTa architecture with the ELECTRA RTD pre-training approach which provided better results than both ELECTRA and DeBERTa.

Recently, we have acquired access to more clinical data, resulting in a total amount of approximately 900,000 EHRs that we can use to pre-train language models. Moreover, we have gained access to more compute power that allows us to pre-train larger language models and to pre-train the models for a longer time. In the future, we would like to pre-train a large DeBERTaV3 model on the new EHR corpus. We expect this architecture to improve our results on the various evaluation datasets that have been established, and that it can solve future clinical problems better than our current models.

Development of domain-adapted clinical language models

MeDa-BERT was developed by initializing it from an existing language model and pre-training it on an in-domain medical dataset. Similarly, we could pre-train a clinical language model from an existing Danish or multilingual language model [12]. Creating a clinical language model in this way might create a superior model compared to pre-training one from scratch because of its increased exposure to data in the pre-training stage. However, the disadvantage of this approach is that the existing model might have encoded biases from its initial pre-training corpus which often stems from online text including debate forums. This creates an increased risk of incorporating biases into the clinical models and, therefore, there is a tradeoff between including text from online sources compared to only including the narrative text of EHRs.

Development of generative pre-trained language models

The main focus of this dissertation has been on developing pre-trained language models using the encoder of the transformer architecture. The reason for this is that the encoder model is specifically trained and constructed to solve classification tasks, which was the main objective of this dissertation. Decoder models, on the other hand, are trained to solve generative tasks such as question answering and text summarization. The performance of generative models such as GPT4 [79] has been surprisingly good, and not only for the typical generative tasks such as text summarization. Generative models have achieved SOTA performance on multiple NLP benchmarks and are able to produce highly coherent and contextually relevant passages [35, 80]. Generative models could be used as assistants to summarize and provide information queried by healthcare professionals. Healthcare professionals could for example ask questions regarding the information in the EHR and the generative model will provide the answer with

references to the relevant medical notes. Moreover, generative models are also capable of using APIs which allows them to create graphs requested by the healthcare professional [81]. However, generative models also have imitations. For example, they can generate content that is misleading or untrue and the same question framed differently can sometimes lead to different responses [82]. Generative models should therefore be tested thoroughly for potential biases. Moreover, the use of commercial models such as GPT-4 is limited for clinical tasks because the sensitive information of EHRs cannot be uploaded to these models. Consequently, we have not been able to use or compare our results with these models. However, open-source generative models with similar performance to GPT-4 have been introduced recently [83, 84]. In the future, we will evaluate and compare the performance of such models for Danish clinical tasks.

Bibliography

[1] *Journalføringsbekendtgørelsen, BEK nr 1225 af 08/06/2021, Bekendtgørelse om autoriserede sundhedspersoners patientjournaler (journalføring, opbevaring, videregivelse, overdragelse m.v.), Indenrigs- og Sundhedsministeriet*. URL: https://www.retsinformation.dk/eli/lta/2021/1225.

[2] Hyoun-Joong Kong. "Managing unstructured big data in healthcare system". In: *Healthcare informatics research* vol. 25, no. 1, 2019, pp. 1–2.

[3] Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. "Quantifying clinical narrative redundancy in an electronic health record". In: *Journal of the American Medical Informatics Association* vol. 17, no. 1, Jan. 2010, pp. 49–53. DOI: 10.1197/jamia.M3390. URL: https://doi.org/10.1197/jamia.M3390.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*, 2018.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners". In: *Advances in neural information processing systems* vol. 33, 2020, pp. 1877–1901.

[6] Beaulieu-Jones Brett K. Kalinin Alexandr A.-Do Brian T. Way Gregory P. Ferrero Enrico Agapow Paul-Michael Zietz Michael Hoffman Michael M. Xie Wei Rosen Gail L. Lengerich Benjamin J. Israeli Johnny Lanchantin Jack Woloszynek Stephen Carpenter Anne E. Shrikumar Avanti Xu Jinbo Cofer Evan M. Lavender Christopher A. Turaga Srinivas C. Alexandari Amr M. Lu Zhiyong Harris David J. DeCaprio Dave Qi Yanjun Kundaje Anshul Peng Yifan Wiley Laura K. Segler Marwin H. S. Boca Simina M. Swamidass S. Joshua Huang Austin Gitter Anthony Ching Travers Himmelstein Daniel S. and Greene Casey S. "Opportunities and obstacles for deep learning in biology and medicine". In: *Journal of The Royal Society Interface* vol. 15, no. 141, 2018, p. 20170387.

[7] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in bioinformatics* vol. 19, no. 6, 2018, pp. 1236–1246.

[8] Oliver Baclic, Matthew Tunis, Kelsey Young, Coraline Doan, Howard Swerdfeger, and Justin Schonfeld. "Challenges and opportunities for public health made possible by advances in natural language processing". In: *Canada communicable disease report = Releve des maladies transmissibles au Canada* vol. 46, no. 6, 2020, 161—168. ISSN: 1188-4169. DOI: 10.14745/ccdr.v46i06a02. URL: https://europepmc.org/articles/PMC7343054.

[9] David E Bloom, Somnath Chatterji, Paul Kowal, Peter Lloyd-Sherlock, Martin McKee, Bernd Rechel, Larry Rosenberg, and James P Smith. "Macroeconomic implications of population ageing and selected policy responses". In: *The Lancet* vol. 385, no. 9968, 2015, pp. 649–657.

[10] Dirk Hovy and Shrimai Prabhumoye. "Five sources of bias in natural language processing". In: *Language and Linguistics Compass* vol. 15, no. 8, 2021, e12432.

[11] Bolette Sandford Pedersen, Sussi Olsen, and Lina Henriksen. *Deliverable D1. 9 Report on the Danish Language*. 2022.

[12] Dan Saattrup Nielsen. "ScandEval: A Benchmark for Scandinavian Natural Language Processing". In: *The 24rd Nordic Conference on Computational Linguistics*. 2023.

[13] Robert Leaman, Ritu Khare, and Zhiyong Lu. "Challenges in clinical natural language processing for automated disorder normalization". In: *Journal of biomedical informatics* vol. 57, 2015, pp. 28–37.

[14] Patrick Ruch, Robert Baud, and Antoine Geissbühler. "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record". In: *Artificial intelligence in medicine* vol. 29, no. 1-2, 2003, pp. 169–184.

[15] Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Cristina Soguero-Ruiz, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, and Pernille Just Vinholt. "Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. © 2022 IEEE. Reprinted, with permission from the authors. 2022, pp. 1–4.

[16] Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Pernille Just Vinholt, and Thiusius Rajeeth Savarimuthu. "MeDa-BERT: A medical Danish pretrained transformer model". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 301–307. URL: https://aclanthology.org/2023.nodalida-1.31.

[17] Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Pernille Just Vinholt, Anne Bryde Alnor, and Thiusius Rajeeth Savarimuthu. "Investigating anatomical bias in clinical machine learning algorithms". In: *Findings of the European Chapter of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023.

[18] Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, Cathrine Brødsgaard Nielsen, Lou-Ann Christensen Andersen, Søren Andreas Just, and Pernille Just Vinholt. "Deep learning detects and visualizes bleeding events in electronic health records". In: *Research and practice in thrombosis and haemostasis* vol. 5, no. 4, 2021, e12505.

[19] Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Rasmus Søgaard Hansen, Thiusius Rajeeth Savarimuthu, Rasmus Bank Lynggaard, and Pernille Just Vinholt. "Doctors

identify hemorrhage better during chart review when assisted by artificial intelligence". English. In: *Applied Clinical Informatics, © Georg Thieme Verlag KG*, July 2023. ISSN: 1869-0327. DOI: 10.1055/a-2121-8380.

[20] Martin Sundahl Laursen, Jannik skyttegaard Pedersen, Rasmus Hansen, Thiusius Rajeeth Savarimuthu, and Pernille Just Vinholt. "Danish Clinical Named Entity Recognition and Relation Extraction". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 655–666. URL: https://aclanthology.org/2023.nodalida-1.65.

[21] Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Pernille Just Vinholt, Rasmus Søgaard Hansen, and Thiusius Rajeeth Savarimuthu. "Benchmark for Evaluation of Danish Clinical Word Embeddings". In: *Northern European Journal of Language Technology* vol. 9, no. 1, 2023.

[22] Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Søren Andreas Just, Thiusius Rajeeth Savarimuthu, Brian Blomholt, Jakob KH Andersen, and Pernille Just Vinholt. "Factors Facilitating the Acceptance of Diagnostic Robots in Healthcare: A Survey". In: *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2022, pp. 442–448.

[23] Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Pernille Just Vinholt, and Thiusius Rajeeth Savarimuthu. "Automatic Annotation of Training Data for Deep Learning Based De-identification of Narrative Clinical Text". In: *WNLPe-Health 2022: Proceedings of The First Workshop on Context-aware NLP in eHealth (WNLPe-Health 2022)*. CEUR Workshop Proceedings. 2023, pp. 30–44.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* vol. 30, 2017.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*, 2013.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* vol. 26, 2013.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[28] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* vol. 5, 2017, pp. 135–146.

[29] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* vol. 9, no. 8, 1997, pp. 1735–1780.

[30] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078*, 2014.

[31] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*, 2014.

[32] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1412–1421.

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[34] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. "Overview of the Transformer-based Models for NLP Tasks". In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2020, pp. 179–183.

[35] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. "GLUE: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461*, 2018.

[36] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250*, 2016.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*, 2019.

[38] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[39] Iz Beltagy, Matthew E Peters, and Arman Cohan. "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150*, 2020.

[40] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942*, 2019.

[41] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. "Efficient Transformers: A Survey". In: *ACM Comput. Surv.* vol. 55, no. 6, 2022. ISSN: 0360-0300. DOI: 10.1145/3530811. URL: https://doi.org/10.1145/3530811.

[42] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets". In: *arXiv preprint arXiv:1906.05474*, 2019.

[43] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. "Don't stop pretraining: Adapt language models to domains and tasks". In: *arXiv preprint arXiv:2004.10964*, 2020.

[44] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. "Publicly Available Clinical BERT Embeddings". In: *NAACL HLT 2019*, 2019, p. 72.

[45] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission". In: *arXiv preprint arXiv:1904.05342*, 2019.

[46] Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. "Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art". In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 2020, pp. 146–157.

[47] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang,

Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. "A large language model for electronic health records". In: *NPJ Digital Medicine* vol. 5, no. 1, 2022, p. 194.

[48] Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. "The lacunae of Danish natural language processing". In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. 2019, pp. 356–362.

[49] Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. "Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text". In: *Journal of the American Medical Informatics Association* vol. 20, no. 5, 2013, pp. 947–953.

[50] Francisco S Roque, Peter B Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Søeby, Søren Bredkjær, Anders Juul, Thomas Werge, Lars J. Jensen, and Søren Brunak. "Using electronic patient records to discover disease correlations and stratify patient cohorts". In: *PLoS computational biology* vol. 7, no. 8, 2011, e1002141.

[51] Alexander Turchin and Luisa F Florez Builes. "Using natural language processing to measure and improve quality of diabetes care: a systematic review". In: *Journal of Diabetes Science and Technology* vol. 15, no. 3, 2021, pp. 553–560.

[52] Cecilia Engel Thomas, Peter Bjødstrup Jensen, Thomas Werge, and Søren Brunak. "Negation scope and spelling variation for text-mining of Danish electronic patient records". In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*. 2014, pp. 64–68.

[53] Kenneth C Enevoldsen, Andreas A Danielsen, Christopher Rohde, Oskar H Jefsen, Kristoffer L Nielbo, and Søren D Østergaard. "Monitoring of COVID-19 pandemic-related psychopathology using machine learning". In: *Acta Neuropsychiatrica* vol. 34, no. 3, 2022, pp. 148–152.

[54] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[55] Benjamin Skov Kaas-Hansen, Davide Placido, Cristina Leal Rodríguez, Hans-Christian Thorsen-Meyer, Simona Gentile, Anna Pors Nielsen, Søren Brunak, Gesche Jürgens, and Stig Ejdrup Andersen. "Language-agnostic pharmacovigilant text mining to elicit side effects from clinical notes and hospital medication records". In: *Basic & clinical pharmacology & toxicology* vol. 131, no. 4, 2022, pp. 282–293.

[56] Batya Friedman and Helen Nissenbaum. "Bias in computer systems". In: *ACM Transactions on information systems (TOIS)* vol. 14, no. 3, 1996, pp. 330–347.

[57] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* vol. 366, no. 6464, 2019, pp. 447–453.

[58] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. "Mitigating Gender Bias in Natural Language Processing: Literature Review". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1630–1640.

[59] Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. "Bias and

fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups". In: *Journal of the American Medical Informatics Association* vol. 28, no. 11, 2021, pp. 2393–2403.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets". In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 25–35.

[61] Hervé Decousus, Victor F Tapson, Jean-François Bergmann, Beng H Chong, James B Froehlich, Ajay K Kakkar, Geno J Merli, Manuel Monreal, Mashio Nakamura, Ricardo Pavanello, Franco Piovella Mario Pini, Frederick A Spencer, Alex C Spyropoulos, Alexander G G Turpie, Rainer B Zotz, Gordon Fitzgerald, and Frederick A Anderson. "Factors at admission associated with bleeding risk in medical patients: findings from the IMPROVE investigators". In: *Chest* vol. 139, no. 1, 2011, pp. 69–79.

[62] Vera E Valkhoff, Preciosa M Coloma, Gwen MC Masclee, Rosa Gini, Francesco Innocenti, Francesco Lapi, Mariam Molokhia, Mees Mosseveld, Malene Schou Nielsson, Martijn Schuemie, Frantz Thiessard, Johan van der Lei, Miriam C J M Sturkenboom, and Gianluca Trifirò. "Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk". In: *Journal of clinical epidemiology* vol. 67, no. 8, 2014, pp. 921–931.

[63] A Amin, S Stemkowski, J Lin, and G Yang. "Thromboprophylaxis rates in US medical centers: success or failure?" In: *Journal of Thrombosis and Haemostasis* vol. 5, no. 8, 2007, pp. 1610–1616.

[64] Alpesh Amin, Stephen Stemkowski, Jay Lin, and Guiping Yang. "Appropriate thromboprophylaxis in hospitalized cancer patients". In: *Clin Adv Hematol Oncol* vol. 6, no. 12, 2008, pp. 910–920.

[65] Jean Paul Rwabihama, Etienne Audureau, Marie Laurent, Lalaina Rakotoarisoa, Marc Jegou, Sofiane Saddedine, Sébastien Krypciak, Stéphane Herbaud, Hind Benzengli, Lauriane Segaux, Esther Guery, Gabin Ambime, Marie-Thérèse Rabus, Jean-Guy Perilliat, Jean-Philippe David, and Elena Paillaud. "Prophylaxis of venous thromboembolism in geriatric settings: a cluster-randomized multicomponent interventional trial". In: *Journal of the American Medical Directors Association* vol. 19, no. 6, 2018, pp. 497–503.

[66] Rumeng Li, Baotian Hu, Feifan Liu, Weisong Liu, Francesca Cunningham, David D McManus, and Hong Yu. "Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach". In: *JMIR medical informatics* vol. 7, no. 1, 2019, e10788.

[67] Maxwell Taggart, Wendy W Chapman, Benjamin A Steinberg, Shane Ruckel, Arianna Pregenzer-Wenzler, Yishuai Du, Jeffrey Ferraro, Brian T Bucher, Donald M Lloyd-Jones, Matthew T Rondina, and Rashmee U. Shah. "Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients". In: *JAMA network open* vol. 1, no. 6, 2018, e183451–e183451.

[68] Avijit Mitra, Bhanu Pratap Singh Rawat, David McManus, Alok Kapoor, and Hong Yu. "Bleeding entity recognition in electronic health records: A comprehensive analysis of end-to-end systems". In: *AMIA Annual Symposium Proceedings*. Vol. 2020. American Medical Informatics Association. 2020, p. 860.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[70] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450*, 2016.

[71] Mike Schuster and Kaisuke Nakajima. "Japanese and korean voice search". In: *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2012, pp. 5149–5152.

[72] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL].

[73] Bethesda (MD): National Library of Medicine (US) PubMed [Internet] [1946] [2023-08-25]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/. In.

[74] Sultan Alrowili and K Vijay-Shanker. "BioM-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA". in: *Proceedings of the 20th workshop on biomedical language processing*. 2021, pp. 221–227.

[75] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* vol. 36, no. 4, 2020, pp. 1234–1240.

[76] Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. "Clinical named entity recognition: Challenges and opportunities". In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE. 2016, pp. 1937–1945.

[77] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "Deberta: Decoding-enhanced bert with disentangled attention". In: *arXiv preprint arXiv:2006.03654*, 2020.

[78] Pengcheng He, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing". In: *arXiv preprint arXiv:2111.09543*, 2021.

[79] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[80] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems". In: *Advances in neural information processing systems* vol. 32, 2019.

[81] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. "Gorilla: Large language model connected with massive apis". In: *arXiv preprint arXiv:2305.15334*, 2023.

[82] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation". In: *ACM Comput. Surv.* vol. 55, no. 12, 2023. ISSN: 0360-0300. DOI: 10.1145/3571730. URL: https://doi.org/10.1145/3571730.

[83] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien

Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*, 2023.

[84] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.