Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration

Betty van Aken[1], **Jens-Michalis Papaioannou**[1], **Manuel Mayrdorfer**[2], **Klemens Budde**[2], **Felix A. Gers**[1] **and Alexander Löser**[1]

[1]Beuth University of Applied Sciences Berlin
[2]Charité Berlin
{bvanaken,michalis.papaioannou,gers,aloeser}@beuth-hochschule.de
{manuel.mayrdorfer,klemens.budde}@charite.de

Abstract

Outcome prediction from clinical text can prevent doctors from overlooking possible risks and help hospitals to plan capacities. We simulate patients at admission time, when decision support can be especially valuable, and contribute a novel *admission to discharge* task with four common outcome prediction targets: Diagnoses at discharge, procedures performed, in-hospital mortality and length-of-stay prediction. The ideal system should infer outcomes based on symptoms, pre-conditions and risk factors of a patient. We evaluate the effectiveness of language models to handle this scenario and propose *clinical outcome pre-training* to integrate knowledge about patient outcomes from multiple public sources. We further present a simple method to incorporate ICD code hierarchy into the models. We show that our approach improves performance on the outcome tasks against several baselines. A detailed analysis reveals further strengths of the model, including transferability, but also weaknesses such as handling of vital values and inconsistencies in the underlying data.

1 Introduction

Clinical professionals make decisions about patients under strong time constraints. The patient information at hand is often unstructured, e.g. in the form of clinical notes written by other medical personnel in limited time. Clinical decision support (CDS) systems can help in these scenarios by pointing towards related cases or certain risks. Clinical outcome prediction is a fundamental task of CDS systems, in which the patient's development is predicted based on data from their Electronic Health Record (EHR). In this work we focus on textual EHR data available at admission time.

Figure 1 shows a sample admission note with highlighted parts that – according to medical doctors – must be considered when evaluating a patient.

Encoding clinical notes with pre-trained language models. Neural models need to extract relevant facts from such notes and learn complex relations between them in order to associate certain clinical outcomes. Pre-trained language models such as BERT (Devlin et al., 2019) have shown to be able to both extract information from noisy text and to capture task-specific relations in an end-to-end fashion (Tenney et al., 2019; van Aken et al., 2019). We thus base our work on these models and pose the following questions:

- Can pre-trained language models learn to predict patient outcomes from their admission information only?
- How can we integrate knowledge about outcomes that doctors gain from medical literature and previous patients?
- How well would these models work in clinical practice? Are they able to interpret common risk factors? Where are they failing?

Simulating patients at admission time. Existing work on text-based outcome prediction focuses on progress notes after a certain time of a patient's hospitalisation (Huang et al., 2019). This is mostly due to a lack of publicly available admission notes and poses some problems: 1) Doctors might miss specific outcome risks early in admission and 2) progress notes already contain information about clinical decisions made on admission time (Boag et al., 2018). We propose to simulate newly arrived patients by extracting admission notes from MIMIC III discharge summaries. We are thus able

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 881–893

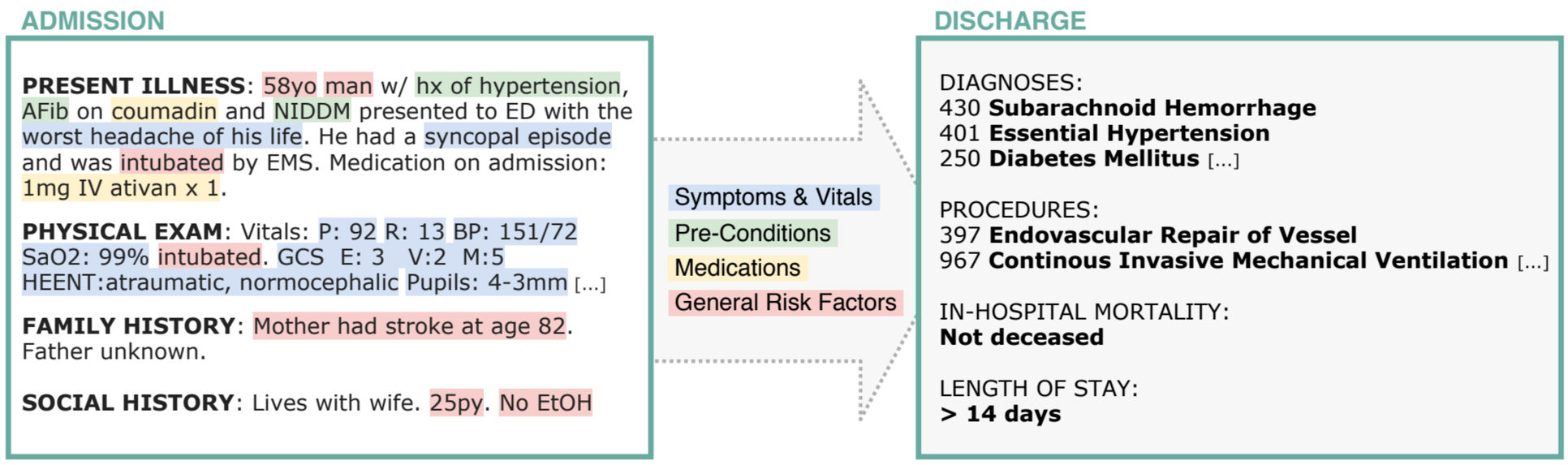

Figure 1: *Admission to discharge* sample that demonstrates the outcome prediction task. The model has to extract patient variables and learn complex relations between them in order to predict the clinical outcome.

to give doctors hints towards possible outcomes from the very beginning of an admission and can potentially prevent early mistakes. We can also help hospitals in planning resources by indicating how long a patient might stay hospitalised.

Integrating knowledge with specialised outcome pre-training. Gururangan et al. (2020) recently emphasized the importance of domain- and task-specific pre-training for deep neural models. Consequently we propose to enhance language models pre-trained on the medical domain with a task-specific *clinical outcome pre-training*. Besides processing clinical language with idiosyncratic and specialized terms, our models are thus able to learn about patient trajectories and symptom-disease associations in a self-supervised manner. We derive this knowledge from two main sources: 1) Previously admitted patients and their outcomes. This knowledge is usually stored by hospitals in unlabelled clinical notes and 2) Scientific case reports and knowledge bases that describe diseases, their presentations in patients and prognoses. We introduce a method for incorporating these sources by creating a suitable pre-training objective from publicly available data.

Contributions. We summarize the major contributions of this work as follows:
1) A novel task setup for clinical outcome prediction that simulates the patient's admission state and predicts the outcome of the current admission.
2) We introduce self-supervised *clinical outcome pre-training*, which integrates knowledge about patient outcomes into existing language models.
3) We further propose a simple method that injects hierarchical signals into ICD code prediction.
4) We compare our approaches against multiple baselines and show that they improve performance on four relevant outcome prediction tasks with up to 1,266 classes. We show that the models are transferable by applying them to a second public dataset without additional fine-tuning.
5) We present a detailed analysis of our model that includes a manual evaluation of samples conducted by medical professionals.

2 Related Work

Using clinical notes for outcome prediction. Boag et al. (2018) studied the predictive value of clinical notes with simple approaches such as bag-of-words. Recent work increasingly applies neural models to compensate for the noisy nature of the data and the complexity of patterns. Hashir and Sawhney (2020) used both convolutional and recurrent layers for outcome prediction, while Jain et al. (2019) and Qiao et al. (2019) proposed attention-based approaches. Dligach et al. (2019) explored pre-training as a strategy to mitigate data sparsity in clinical setups. Si and Roberts (2019) and Suresh et al. (2018) further showed that outcome prediction benefits from a multitask setup. In contrast to earlier work we apply neural models to admission notes in an *admission to discharge* setup.

Pre-trained language models for the clinical domain. While pre-trained language models are successful in many areas of NLP, there has been little work on applying them to the clinical domain (Qiu et al., 2020). Alsentzer et al. (2019) and Huang et al. (2019) both pre-trained BERT-based models on clinical data. They evaluated their work on readmission prediction and other NLP tasks. We are the first to evaluate pre-trained language models on multiple clinical outcome tasks with large label sets. We further propose a novel pre-training objective specifically for the clinical domain.

Prediction of diagnoses and procedures. The majority of work on diagnosis and procedure prediction covers either single diagnoses (Liu et al., 2018; Choi et al., 2018) or coarse-grained groups (Peng et al., 2020; Sushil et al., 2018). We argue that models should predict diseases and procedures in a fine-grained manner to be beneficial for doctors. Thus we use all diagnosis and procedure codes from the data for our outcome prediction tasks.

ICD coding vs. outcome prediction. There is a variety of work in the related field of automated ICD coding (Xie et al., 2018; Falis et al., 2019). Zhang et al. (2020) recently presented a model able to identify up to 2,292 ICD codes from text. However, ICD coding differs from outcome prediction in the way that diseases are directly extracted from text rather than inferred from symptom descriptions and patient history. We further discuss this distinction in Section 6.

3 Clinical *Admission to Discharge* Task

Clinical outcome prediction can be defined in different ways. We approach the task from a doctor's perspective and predict the outcome of a current admission from the time of the patient's arrival to the hospital unit. We describe our setup as follows.

3.1 Clinical Notes from MIMIC III

As our primary data source, we use the freely-available MIMIC III v1.4 database (Johnson et al., 2016). It contains de-identified EHR data including clinical notes in English from the Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center in Massachusetts between 2001 and 2012. We focus our work on discharge summaries in particular and the outcome information associated with an admission. Similar to previous work, we filter out notes about newborns and remove duplicates.

3.2 Creating Admission Notes from Discharge Summaries

The state of a patient is commonly summarized in an ongoing document, which finally concludes in a discharge summary. Since we want to support clinical decisions from the beginning of a patient's stay, we simulate the state of the patient's document at admission time. We thus filter the document by sections that are known at admission such as: *Chief complaint, (History of) Present illness, Medical history, Admission Medications, Allergies, Physical exam, Family history* and *Social history*. We further describe the filtering in Appendix B.1. Our approach results in 48,745 admission notes. As shown in Table 1 the notes contain about 400 words on average. The selection of admission sections as well as the resulting structure of the notes were verified by medical doctors.

This newly created admission dataset enables us to make predictions on the outcome of a current admission. At inference time, doctors can then use the model's predictions on textual data from newly arrived patients.

Admission Notes Statistics			
avg (words / doc)	std (words / doc)	avg (sent / doc)	std (sent / doc)
396.3	233.3	32.5	23.1

Table 1: Numbers of words / sentences in MIMIC III admission notes. We see a high variation in length.

Multi-label tasks: ICD-9 codes per dataset split							
Diagnoses				Procedures			
Total	Train	Val	Test	**Total**	Train	Val	Test
1,266	1,201	906	1,031	**711**	672	476	563

Table 2: Distribution of ICD-9 codes per dataset split (patient-wise). Note that very rare codes do not appear in each split of the dataset.

Single-label tasks: Samples per class					
Mortality		Length of Stay (in days)			
0	1	≤ 3	> 3 & ≤ 7	> 7 & ≤ 14	> 14
43,609	5,136	5,596	16,134	13,391	8,488

Table 3: Distribution of labels for *Mortality Prediction* and *Length of Stay* task. Both tasks have unbalanced class distributions.

3.3 Outcome Prediction Tasks

We select four relevant tasks for outcome prediction in consultation with medical professionals. All tasks take admission notes as input.

Diagnosis prediction. A main goal of clinical outcome prediction is to support medical professionals in the process of differential diagnosis. We thus take all diagnoses associated with an admission into account and frame the task as an extreme multi-label classification. Diagnoses are encoded as ICD-9 codes in the MIMIC III database. Following Choi et al. (2017), we group ICD-9 diagnosis codes from the database from 4- into 3-digit codes to reduce complexity while still obtaining granular suggestions. This results in a total of 1,266 diag-

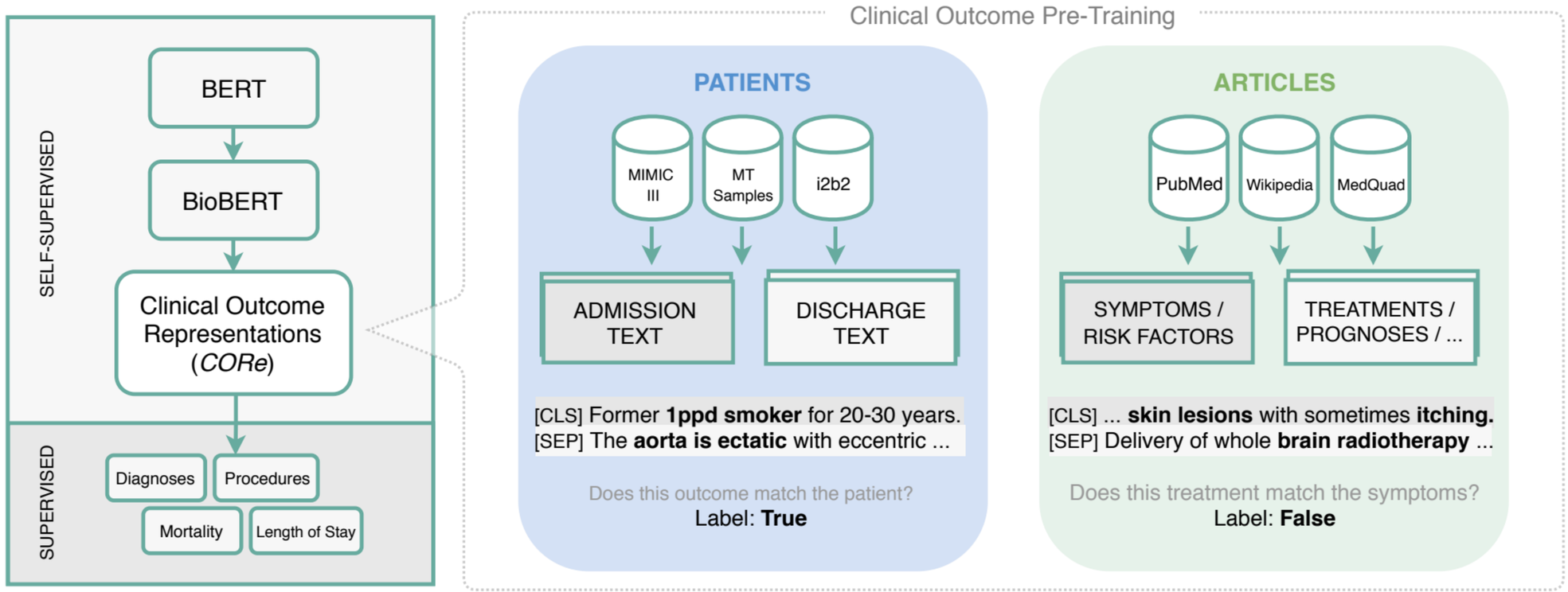

Figure 2: Schematic demonstration of *clinical outcome pre-training*. Sources of clinical knowledge are complete patient notes and medical articles. Based on that we create a self-supervised learning objective that teaches relations between symptoms, risk factors and outcomes.

nosis codes, which are distributed over our dataset splits as shown in Table 2. The labels are power-law distributed with a long tail of very rare codes.

Procedure prediction. Procedures are either diagnostics or treatments applied to a patient during a stay. Similarly to diagnosis prediction, this is an extreme multi-label task. We again group the ICD-9 codes from the MIMIC III database into 3-digit codes. In total there are 711 procedure codes labelled in the database in a power law distribution similar to the diagnosis codes.

In-hospital mortality prediction. Predicting a patient's mortality risk is a fundamental part of the triage process. In-hospital mortality in particular describes whether a patient died during the current admission and is a binary classification task. The percentage of deceased patients in the data is around 10% (see Table 3). As some notes contain direct indications of mortality such as *patient deceased* within the admission sections, we apply an additional filter for those terms.

Length-of-stay prediction. The duration of an ICU stay is an important information for hospitals in order to plan allocations of resources. We group patients into four major categories regarding their length of stay: *Under 3 days, 3 to 7 days, 1 week to 2 weeks, more than 2 weeks.* These categories were recommended by medical doctors in order to make the results as useful as possible in clinical practice. Table 3 shows the samples per class.

4 Integrating Clinical Knowledge Into Language Models

We propose *clinical outcome pre-training*, a way to integrate knowledge about clinical outcomes into pre-trained language models. We further introduce an additional step to incorporate ICD code hierarchy into our multi-label classification tasks.[1]

4.1 Clinical Outcome Pre-Training

Motivation. Language model pre-training has shown to be of use in specialised domains like the clinical (Alsentzer et al., 2019; Huang et al., 2019). However, these models lack knowledge about patient trajectories and symptom-diagnosis relations, because their training is focused on learning language characteristics.

We develop an additional pre-training step that produces *Clinical Outcome Representations (CORe)* in order to teach the model relations between symptoms, risk factors and clinical outcomes. Much of this knowledge is present and publicly available, e.g. in knowledge bases like Wikipedia or publication archives like PubMed. Another source is available to hospitals in the form of unlabelled clinical notes from previous patients. The suggested outcome pre-training is a way to use this knowledge to improve the model's capabilities in predicting clinical outcomes as described in 3.3.

Corresponding to the way doctors gain their knowledge from both experience and medical literature,

[1]The code to recreate the experiments and datasets described in this paper is accessible at: https://github.com/bvanaken/clinical-outcome-prediction

we incorporate knowledge from complete patient notes (including discharge information) and medical articles.

Training objective. Our proposed training objective (Figure 2) is strongly related to the Next Sentence Prediction (NSP) task introduced by Devlin et al. (2019). In NSP the model gets two sentences as an input and predicts whether the second follows the first sentence. This way models such as BERT learn relations between sentences. We convert this setting so that the model instead learns relations between admissions and outcomes.
From common sections in patient notes, we create two categories: Sections that are created at admission A and sections that are created after admission, e.g. at discharge time D. Given a patient note N, we split it into sections $A_N \in A$ and $D_N \in D$. We remove all other sections. We then sample token sequences from these sections to get $t_{N,1...k} \in A_N$ and $t'_{N,1...k} \in D_N$, where k is randomly set between 30 and 50 tokens. We then train the model to maximize $P(Same\_Patient|X_{N\_N})$ and $P(Other\_Patient|X_{N\_M})$ with

$$X_{N\_N} = Enc(t_{N,1...k}, t'_{\mathbf{N},1...k})$$
$$X_{N\_M} = Enc(t_{N,1...k}, t'_{\mathbf{M},1...k}) \quad (1)$$

with M being a randomly sampled document from the same batch and Enc referring to the BioBERT encoding. As in the original NSP setting, we apply negative sampling (X_{N_M}) for 50% of examples. We apply the same strategy on medical articles and case reports, so that A represents sections describing symptoms and risk factors, and D represents sections that describe outcomes of a disease or case.

Data sources. We create the pre-training dataset from multiple public sources. To integrate knowledge that doctors gain from previous patients and medical literature, we create two groups of sources:
1) *Patients*, which includes 32,721 discharge summaries from the MIMIC III training set, 5,000 publicly available medical transcriptions from the MTSamples website [2] and 4,777 clinical notes from the i2b2 challenges 2006-2012[3] (Uzuner et al., 2007, 2008, 2010a,b, 2011, 2012; Sun et al., 2013b,a).
2) *Articles*, composed of 9,335 case reports from PubMed Central (PMC), 2,632 articles from Wikipedia describing diseases and 1,467 article sections from the MedQuAd dataset (Abacha and Demner-Fushman, 2019) extracted from NIH websites such as cancer.gov.
While *Patients* samples contain unaudited practical knowledge, *Articles* samples are built from verified general medical knowledge such as peer-reviewed studies. The sources are therefore substantially different and we evaluate their individual effect on performance in Section 5.3.

Data preparation. We create admission (A_N) and discharge parts (D_N) of the documents based on section headings. We define common sections belonging to the admission part and those belonging to the discharge part similar to the method described in Section 3.2. We ignore sections that cannot be categorized. For section heading extraction from MIMIC III discharge summaries and MTSamples transcriptions, we apply simple rule-based approaches, which is feasible because the notes are well-structured. For Wikipedia we use headings from the WikiSection dataset (Arnold et al., 2019) filtered for disease articles only. For PubMed Central we similarly use the PubMedSection dataset (Schneider et al., 2020) and filter for section headings that indicate case reports. As i2b2 notes are less well-structured in comparison to MIMIC III discharge summaries, we use a classifier as proposed by Rosenthal et al. (2019) to determine which section a sentence belongs to. The classifier is trained on an annotated set of i2b2 notes and then applied to all other notes.

4.2 ICD+: Incorporation of ICD Hierarchy

Medical knowledge in ICD labels. Diagnosis and procedure prediction requires the model to predict ICD-9 codes in a multi-label manner. ICD-9 codes are hierarchically ordered into associated groups. Figure 3 shows the code hierarchy for *Malignant hypertensive renal disease* with the ICD-9 code *403.0*. The diagnosis has two parent groups namely *Hypertension renal disease* and *Diseases of the circulatory system*. Diagnoses or procedures in the same group often share similar medical characteristics, therefore hierarchical relations of a labelled code can be valuable information. This medical information is currently not integrated into the model. The same holds for words describing the ICD-9 codes, that often represent further important signals, such as the words *renal* or *malignant*.

[2] https://mtsamples.com

[3] We exclude notes from the 2014 De-identification and Heart Disease Risk Factors Challenge in order to use this set for evaluation as described in Section 5.4.

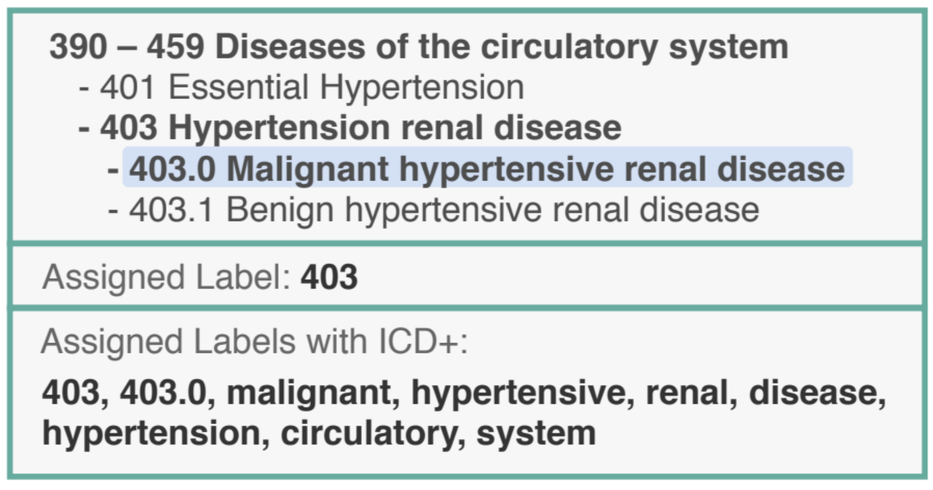

Figure 3: Example of *ICD+* labelling. *Malignant hypertensive renal disease* is assigned to nine codes (bottom row) that inform about the type and group of the disease.

Enhancing training with useful additional signals. We propose a simple method, *ICD+*, to incorporate both associated groups and words into the model weights: Instead of only classifying 3-digit codes (as mentioned in 3.3), we let the model additionally predict the 4-digit codes and the bag of associated words with a code and its parent groups. In order to create the bag of words per code, we use the descriptions of ICD-9 codes from MIMIC III and remove all stop words. As shown in Figure 3, the *ICD+* method assigns eight additional labels to the example diagnosis and therefore supplies the model with further information about the diagnosis during training.
By increasing the amount of labels per sample, we integrate relevant medical knowledge and enable the model to learn implicit relations between codes and code groups that share certain words. We evaluate the effectiveness of *ICD+* in Section 5.

5 Experimental Evaluation

5.1 Training Clinical Outcome Representations

We pre-train the *CORe* model on top of BioBERT weights[4]. We then fine-tune the model separately on the four outcome tasks. We use the same training regimen for both pre-training and fine-tuning: We tokenize the texts with WordPiece tokenization and truncate them to 512 tokens, due to the limited context length of the pre-trained models. We use early stopping and tune hyperparameters as described in Appendix C.

[4]We choose BioBERT as the base for our model because it outperforms BERT on medical tasks and has not seen data from our test set during pre-training unlike DischargeBERT.

5.2 Baseline Models

In the following, we introduce the baseline models that we evaluate on the novel outcome prediction tasks. In order to understand the abilities of pre-trained language models we compare their performance against more traditional approaches. The first three models (*BOW, word embeddings, CNN*) are trained using the hyperparameters proposed by the authors for outcome prediction tasks. The language models are fine-tuned the same way as the *CORe* model.

Bag-of-Words. Boag et al. (2018) shows that a simple bag-of-words (BOW) approach can outperform more complex models on tasks like mortality prediction. We thus include their approach in our evaluation. We adopt their training setting except that we consider 200 instead of 20 top tf-idf words in order to make the model converge.

Pre-trained word embeddings. Boag et al. (2018) further propose the use of pre-computed word embeddings that were trained on MIMIC III data. We use the same setting as for the BOW approach and fit a support vector machine classifier on the clinical outcome tasks.

Convolutional Neural Network (CNN). Si and Roberts (2019) built a neural network for mortality prediction with two hierarchical convolutional layers at the word and sentence levels and then aggregated it to a patient level representation. We follow their approach to evaluate the model on our four *admission to discharge* tasks.

BioBERT. Following the success of BERT, Lee et al. (2020) further pre-trained the model on biomedical research articles from PubMed using abstracts and full-text articles. They reported improved performance on a range of biomedical text mining tasks.

ClinicalBERT and DischargeBERT. We further evaluate two public language models pre-trained on the clinical domain, with MIMIC III data in particular. Huang et al. (2019) pre-trained a BERT Base model on 100,000 random clinical notes (ClinicalBERT) while Alsentzer et al. (2019) further pre-trained BioBERT on all discharge summaries from MIMIC III (we refer to the model as DischargeBERT for simplicity).

	Diagnoses (1266 classes)	Procedures (711 classes)	In-Hospital Mortality (2 classes)	Length-of-Stay (4 classes)
BOW (Boag et al., 2018)	75.87	77.47	79.15	65.83
Embeddings (Boag et al., 2018)	75.16	76.72	79.94	66.78
CNN (Si and Roberts, 2019)	61.18	73.13	75.50	64.49
BERT Base (Devlin et al., 2019)	82.08	85.84	81.13	70.40
ClinicalBERT (Huang et al., 2019)	81.99	86.15	82.20	71.14
DischargeBERT (Alsentzer et al., 2019)	*82.86*	*87.09*	***84.51***	*71.73*
BioBERT Base (Lee et al., 2020)	82.81	86.36	82.55	71.59
BioBERT ICD+	83.17	87.45	-	-
CORe Articles (w/o ICD+)	83.46 (82.89)	87.43 (86.75)	83.64	71.99
CORe Patients (w/o ICD+)	83.41 (83.40)	**88.37** (86.60)	83.60	71.96
CORe All (w/o ICD+)	**83.54** (83.39)	87.65 (87.15)	**84.04**	**72.53**

Table 4: Results on outcome prediction tasks in macro-averaged % AUROC. The *CORe* models outperform the baselines, *ICD+* adds further improvement (values in parentheses are ablation results without *ICD+*). DischargeBERT results are printed in italic because the model has seen all test data during pre-training and is therefore slightly advantaged.

5.3 Results on MIMIC III Admission Notes

Table 4 shows performances in (macro-averaged) area under the receiver operating characteristic curve (AUROC). We report scores of the *CORe* model trained only on *Articles*, *Patients* and in a combined training setting *CORe All*. We evaluate diagnosis and procedure prediction both with and without the *ICD+* method on BioBERT and the *CORe* models. In both scenarios we evaluate on 3-digit ICD codes only, in order to maintain comparability between the methods.

Pre-trained models outperform baselines. We see that the evaluated pre-trained language models clearly outperform the *BOW*, *word embeddings* and *CNN* approaches. We further observe that the *CORe* models improve scores on all tasks in comparison to the baseline models, except for DischargeBERT that reaches a higher score in mortality prediction – probably affected by its exposure to the test data. This shows that even though the language models are trained on similar data (e.g. PubMed and/or clinical notes), the specific *outcome pre-training* improves the model's ability to predict clinical outcome targets. Pre-training on *Patients* and *Articles* achieve similar improvements over the baselines, while the combined training is the most effective. An exception is the procedure prediction, where pre-training on *Patients* achieves the highest score. A probable reason is that procedures are documented in more detail in clinical notes, especially since our selection of medical articles focuses on diseases rather than procedures.

Predicting mortality risk is easier than length of stay. We see that the models reach higher scores in the binary mortality task than in length of stay prediction. Even a simple *BOW* approach can reach a relatively high score, which indicates that most of the notes contain clear hints towards an increased mortality risk. On the other hand, the length of stay task is difficult due to the many factors that can contribute to the length of a patient's stay after the admission, including nonclinical factors such as the patient's insurance situation (Khosravizadeh et al., 2016).

ICD hierarchy improves diagnosis and procedure predictions. Table 4 shows an ablation test without the *ICD+* method (in parentheses). We see that both the BioBERT model and the *CORe* models improve when incorporating code hierarchy and relations through *ICD+* into the training process. This is especially visible for ICD procedures, where the hierarchical and textual information, e.g. that a *Nephropexy* is an *operation* on the *kidney* can add important signals during training.

	i2b2 Diagnoses
BioBERT ICD+	80.43
CORe Articles	81.46
CORe Patients	**82.31**
CORe All	81.15

Table 5: Results on i2b2 diagnosis prediction task (5 classes) in % AUROC. The models reach similar results as on the MIMIC III data, indicating their transferability to other data sources without additional fine-tuning.

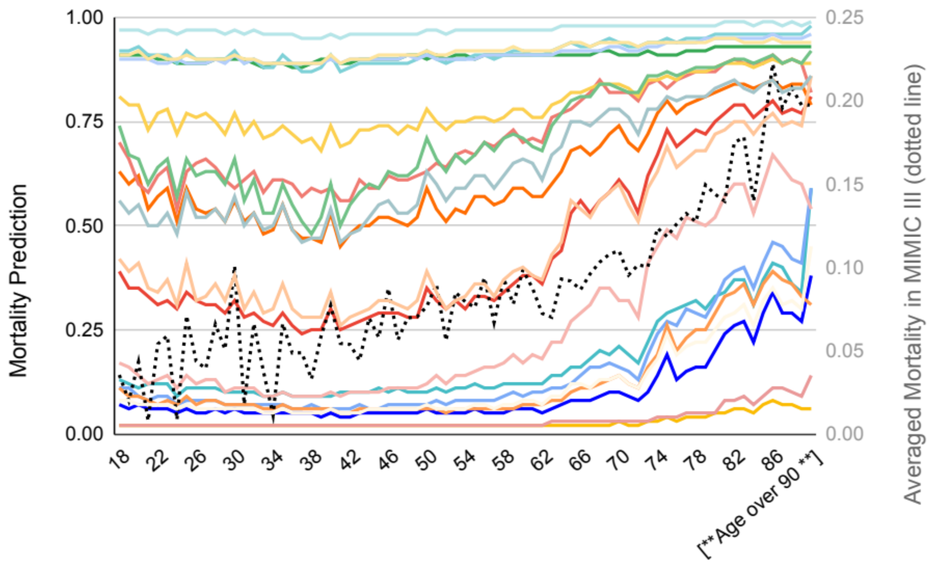

Figure 4: Impact of age on mortality prediction on 20 random samples. Mortality risk and age mostly increase proportionally as intended, with certain peaks that might indicate unintended biases in the data.

5.4 Model Transferability: Cross-Verification on i2b2 Clinical Notes

In order to verify that the fine-tuned models are transferable to ICU data from other sources, we apply it to data from the i2b2 De-identification and Heart Disease Risk Factors Challenge (Stubbs et al., 2015). We convert the clinical notes to admission notes as further described in Appendix B.2, which results in 1,118 samples labelled with up to five ICD-9 codes.

Models generalize to i2b2 data. We apply our MIMIC III-based models to predict diagnosis codes for the i2b2 notes without further fine-tuning. We then evaluate based on whether the predictions contain the five mentioned ICD-9 codes. The results in macro-averaged % AUROC are shown in Table 5. Even though the clinical notes differ from the MIMIC III notes in structure and writing style, the tested models are mostly able to identify the conditions. The scores are comparable to the MIMIC III results, which shows that the models are able to generalise on data from different sources such as other hospitals.

6 Discussion and Findings

Clinical outcome prediction is a sensitive task. We therefore conduct an extensive analysis on the *CORe All* model including a manual error analysis by medical doctors on 20 randomly chosen samples to understand how the model would perform in clinical practice.[5]

[5]Our demo application used for this analysis is available at: https://outcome-prediction.demo.datexis.com

	% AUROC
All Diagnoses	83.54
Diagnoses **Mentioned** in Text	**87.10**
Diagnoses **Not Mentioned** in Text	82.35

Table 6: Analysis of the impact of directly mentioned diagnoses on the diagnosis prediction task. Mentioned diagnoses are detected more reliably. Though on unmentioned diagnoses, scores only see a small decrease compared to the overall score.

6.1 A Closer Look at the Model's Abilities

Does the model mainly extract already present diagnoses? We observe that a majority of coded diseases are already mentioned in the admission text. This is mainly due to chronic diseases (e.g. *diabetes mellitus*) or to conditions that were identified prior to the ICU admission (e.g. in the emergency ward). We want to know if our model is also able to predict diagnoses that are not mentioned in the text. We annotate the admission texts with ICD-9 diagnosis codes with the methodology described by Searle et al. (2020). We then evaluate on codes that were explicitly mentioned in the text and those that were not. Table 6 shows that the model indeed extracts many diagnoses directly from the text and thus reaches a higher score on mentioned diagnoses. On the other hand, we see that the performance on non-mentioned diagnoses does drop only slightly, indicating that the model has also learned to predict non-mentioned diagnoses.

How does age and gender impact predictions? Age and gender are common risk factors with significant impact on the potential clinical outcome of a patient. We want our models to learn that impact without overestimating it. We test the model's behaviour by switching age and gender throughout 20 random samples and analyse how the mortality prediction changes. For each sample we manually switch the age mention and iterate over it from 18 until [**Age over 90**][6]. Figure 4 shows that the analysed samples show a high variation in mortality risk and that age only impacts the prediction partially. In all cases the prediction increases with age – as expected from a medical perspective. We also observe some peaks without a medical reason that are caused by the mortality of certain age groups in the original data (black dotted line). This demonstrates how the model does not follow medical reasoning but merely statistic observations. We

[6]De-identified age information for patients older than 89.

similarly switch the gender mention and all pronouns in the texts and observe that mortality prediction for male patients is increased by 5% on average, consistent with medical rationale.

Where is the model failing?

1. **Negation**: While our error analysis depicts that negation does not generally falsify the model's predictions, we find single samples in which especially medical-specific negations, such as *abstinent from alcohol*, are misinterpreted by the model, e.g. into *alcohol dependence syndrome*.

2. **Numerical data**: Wallace et al. (2019) show BERT's inabilities to interpret numbers. We observe this in the case that the model does not interpret life-threatening vital values (such as temperature over 105°F) as an increased mortality risk. Clinical notes contain many such relevant values, thus improving the encoding of such data is an important goal for future work.

6.2 There is no Ground Truth in Clinical Data

Incomplete and inconsistent labels. Our error analysis reveals that 60% of the analysed samples are partially under-coded. They contain indicators for a diagnosis or procedure but miss the corresponding ICD-9 code. This is consistent with results from Searle et al. (2020) showing that MIMIC III is up to 35% under-coded. Additionally we find that procedures that are almost always performed in the ICU such as *Puncture of vessel* are often coded inconsistently. While a doctor can infer these labels with medical common sense, they pose a challenge to our models. We therefore suggest a critical view towards the data and welcome additional clinical datasets to compensate for noisy labels.

Multiple possible outcomes. 85% of analysed samples contain false positive predictions that the doctors still consider medically reasonable. This demonstrates that there are many possible clinical pathways and that some might not be foreseeable at admission time. We also see many cases in which the information in the clinical note is not sufficient and therefore allows multiple interpretations. For future work, we propose including further EHR data as suggested by Khadanga et al. (2019) to extend the patient representation in these scenarios.

7 Conclusion

We reframe the task of clinical outcome prediction to consider the admission state of a patient and support doctors in their initial decision process. We show that current state-of-the-art language models outperform selected baselines on this task and present methods for further improvement: *Outcome pre-training* enables our models to learn from unlabelled sources and *ICD+* incorporates hierarchical and textual ICD representations into our models. For future work, we suggest considering pre-trained language models with larger context sizes (Beltagy et al., 2020; Zaheer et al., 2020) and languages other than English (Reys et al., 2020). We further encourage work on semantic encoding of negated terms and numerical data from clinical text.

Acknowledgments

We would like to thank Anjali Grover and Sebastian Herrmann for their support throughout the project. Our work is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreement 01MD19003B (PLASS) and 01MK2008MD (Servicemeister).

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511:1–511:23.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pages 1823–1832, Beijing, China. ACM.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. ACL.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics, TACL*, 7:169–184.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *Computing Research Repository*, arXiv/2004.05150.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a Note? Unpacking Predictive Value in Clinical Note Representations. *AMIA Summits on Translational Science Proceedings*, 2018:26 – 34.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305, Boston, Massachusetts. PMLR.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 4552–4562, Montréal, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. ACL.

Dmitriy Dligach, Majid Afshar, and Timothy A. Miller. 2019. Toward a clinical text encoder: pretraining for clinical natural language processing with applications to substance misuse. *J. Am. Medical Informatics Assoc.*, 26(11):1272–1278.

Matús Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios A. Tsaftaris, and Alison O'Neil. 2019. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis, LOUHI@EMNLP 2019*, pages 168–177, Hong Kong, China. ACL.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8342–8360, Online. ACL.

Mohammad Hashir and Rapinder Sawhney. 2020. Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics*, 108:103489.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In *Proceedings of ACM Conference on Health, Inference, and Learning, CHIL 2020*, Online. ACM.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An Analysis of Attention over Clinical Notes for Predictive Tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA. ACL.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Swaraj Khadanga, Karan Aggarwal, Shafiq R. Joty, and Jaideep Srivastava. 2019. Using Clinical Notes with Time Series Data for ICU Management. 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 2019*, pages 6431–6436, Hong Kong, China. ACL.

Omid Khosravizadeh, Soudabeh Vatankhah, Peivand Bastani, Rohollah Kalhor, Samira Alirezaei, and Farzane Doosty. 2016. Factors affecting length of stay in teaching hospitals of a middle-income country. *Electronic physician*, 8(10):3042–3047.

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.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep EHR: Chronic Disease Prediction Using Medical Notes. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2018*, volume 85 of *Proceedings of Machine Learning Research*, pages 440–464, Palo Alto, California, USA. PMLR.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. 2020. Self-Attention Enhanced Patient Journey Understanding in Healthcare System. In *Proceedings of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2020*, Online.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. MNN: Multimodal Attentional Neural Networks for Diagnosis Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5937–5943, Macao, China.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained

Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, 63:1872 – 1897.

Arthur D. Reys, Danilo Silva, Daniel Severo, Saulo Pedro, Marcia M. de Souza e Sá, and Guilherme A. C. Salgado. 2020. Predicting Multiple ICD-10 Codes from Brazilian-Portuguese Clinical Notes. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, Rio Grande, Brazil.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging Medical Literature for Section Prediction in Electronic Health Records. 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 4864–4873, Hong Kong, China. ACL.

Rudolf Schneider, Tom Oberhauser, Paul Grundmann, Felix Alexander Gers, Alexander Loeser, and Steffen Staab. 2020. Is Language Modeling Enough? Evaluating Effective Embedding Combinations. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4739–4748, Marseille, France. ELRA.

Thomas Searle, Zina M. Ibrahim, and Richard J. B. Dobson. 2020. Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020*, pages 76–85. ACL.

Yuqi Si and Kirk Roberts. 2019. Deep Patient Representation of Clinical Notes via Multi-Task Learning for Mortality Prediction. *AMIA Summits on Translational Science Proceedings*, 2019:779–788.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58:S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(6):S5–S12.

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

Harini Suresh, Jen J. Gong, and John V. Guttag. 2018. Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 802–810, London, UK. ACM.

Madhumita Sushil, Simon Šuster, Kim Luyckx, and Walter Daelemans. 2018. Patient representation learning and interpretable evaluation using clinical notes. *Journal of Biomedical Informatics*, 84:103 – 113.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA.

Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R. South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac S. Kohane. 2008. Viewpoint Paper: Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Viewpoint Paper: 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. 2010a. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010b. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

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.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. 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 2019*, pages 5306–5314, Hong Kong, China. ACL.

Pengtao Xie, Haoran Shi, Ming Zhang, and Eric P. Xing. 2018. A Neural Architecture for Automated

ICD Coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 1066–1076, Melbourne, Australia. ACL.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 24–34. Association for Computational Linguistics.

A Distribution of Diagnosis and Procedure Labels

Figure 5 and Figure 6 show the distributions of labels in the diagnosis and procedure prediction training sets. Both distributions follow the power law with a long tail of rare codes.

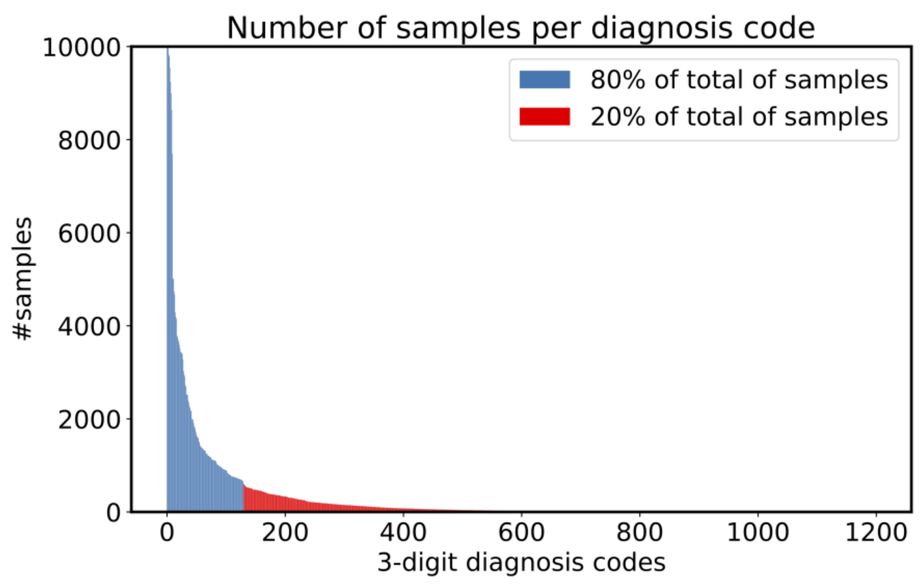

Figure 5: Distribution of ICD-9 diagnosis codes in MIMIC III training set.

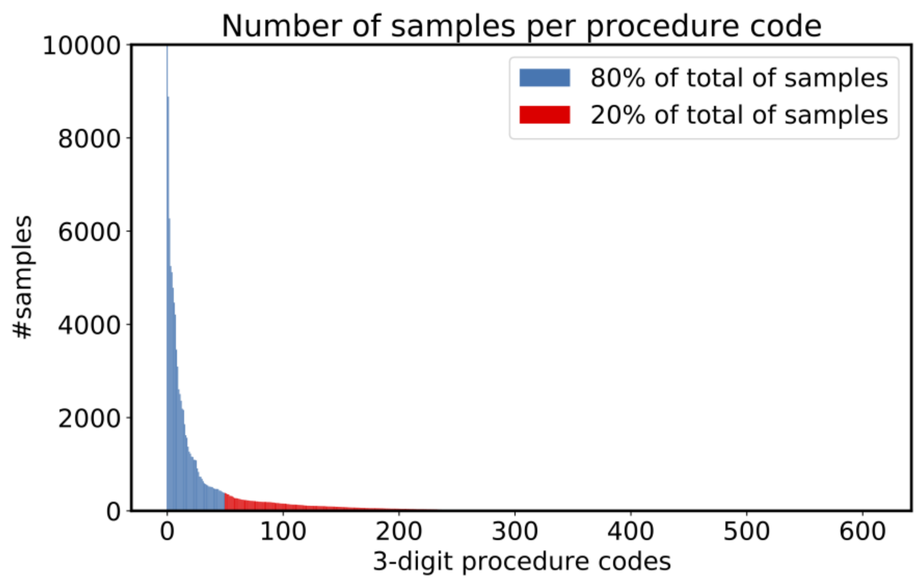

Figure 6: Distribution of ICD-9 procedure codes in MIMIC III training set.

B Pre-Processing Clinical Notes

B.1 Admission Notes From Discharge Summaries

We use MIMIC III discharge summaries that contain aggregated information about a patient such as doctor's assessments, relevant lab values, medications, and the patient's history. In order to filter the documents by admission sections, we first split all discharge summaries into sections with simple pattern matching. Together with clinical professionals, we then evaluated discharge summaries and identified sections that are known at admission time. We remove all other sections and thus hide information about the further hospital course and discharge of a patient. We exclude notes that do not contain any of the admission sections. We further apply a patient-wise split into train, validation and test set with a 70/10/20 ratio.

B.2 Converting i2b2 Data into Admission Discharge Task

The i2b2 De-identification and Heart Disease Risk Factors Challenge (Stubbs et al., 2015; Stubbs and Uzuner, 2015) introduced a dataset that contains clinical notes and discharge summaries annotated based on risk factors and disease indicators. We convert the data into an *admission to discharge* task by selecting five of the annotated conditions which correspond to ICD-9 codes as our labels, namely Hypertension (401), Hyperlipidemia (272), Coronary artery disease (414), Diabetes mellitus (250) and Obesity (278). Just like the MIMIC III diagnosis task, samples are annotated in a multi-label fashion. In order to convert the clinical notes to admission notes, we use the dataset from Rosenthal et al. (2019) that contain section labels per sentence. We then exclude sections that are not known at admission time concurrent to Section 3.2.

C Hyperparameter Setting

We use the following setting for pre-training and fine-tuning of the introduced Transformer-based models: We use early stopping and apply a random search for tuning the following hyperparameters on the validation set: learning rate [1e-4−1e-6], warmup steps [50−30k], dropout [0.1−0.3], class balancing [True/False] (fine-tuning only), gradient accumulation [1−200] with a batch size of 20.

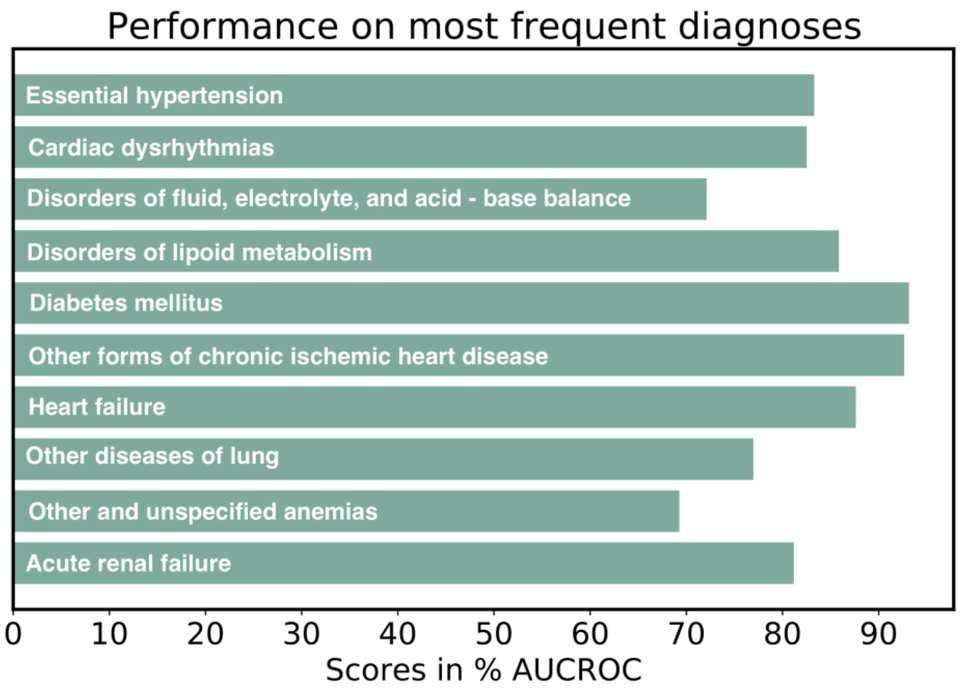

Figure 7: Top 10 diagnoses by frequency with the scores reached by the *CORe All* model.

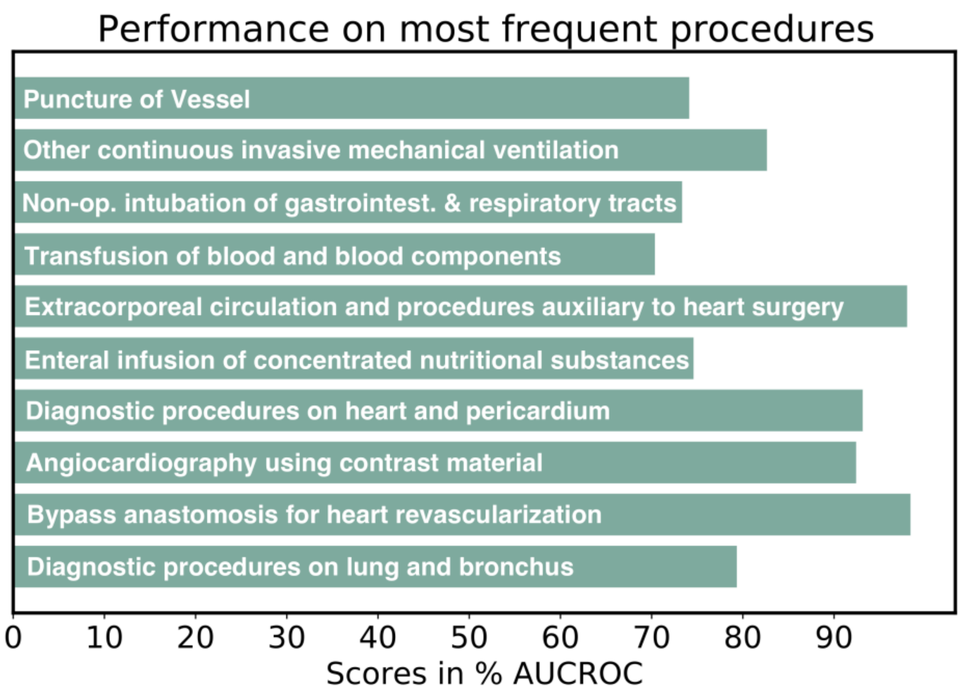

Figure 8: Top 10 procedures by frequency with the scores reached by the *CORe All* model.

D Results on Top 10 Diagnoses and Procedures

Figures 7 and 8 show the % AUROC scores of our *CORe All* model on the most frequent labels within the diagnosis and procedure prediction tasks. Figure 7 show that many chronic diseases such as *Essential Hypertension* or *Chronic ischemic heart disease* are among the most common within the MIMIC III dataset and present with relatively high AUROC values. We also observe that very specific codes such as *Diabetes mellitus* and *Bypass Anastomosis* are predicted more easily compared to more general codes such as *Other and unspecified anemias*.

Figure 8 further shows the negative influence of inconsistent labeling on standard procedures such as *Puncture of Vessel*.