Data Drift in Clinical Outcome Prediction from Admission Notes

Paul Grundmann*, **Jens-Michalis Papaioannou***, **Tom Oberhauser***, **Thomas Steffek***, **Amy Siu***, **Wolfgang Nejdl**†, **Alexander Löser***

*Berliner Hochschule für Technik (BHT) - Luxemburger Straße 10, 13467 Berlin
pgrundmann, toberhauser, mpapaioannou, tsteffek, asiu, aloeser@bht-berlin.de
†Leibniz University Hannover - Welfengarten 1, 30167 Hannover
nejdl@L3S.de

Abstract

Clinical NLP research faces a scarcity of publicly available datasets due to privacy concerns. MIMIC-III marked a significant milestone, enabling substantial progress, and now, with MIMIC-IV, the dataset has expanded significantly, offering a broader scope. In this paper, we focus on the task of predicting clinical outcomes from clinical text. This is crucial in modern healthcare, aiding in preventive care, differential diagnosis, and capacity planning. We introduce a novel clinical outcome prediction dataset derived from MIMIC-IV. Furthermore, we provide initial insights into the performance of models trained on MIMIC-III when applied to our new dataset, with specific attention to potential data drift. We investigate challenges tied to evolving documentation standards and changing codes in the International Classification of Diseases (ICD) taxonomy, such as the transition from ICD-9 to ICD-10. We also explore variations in clinical text across different hospital wards. Our study aims to probe the robustness and generalization of clinical outcome prediction models, contributing to the ongoing advancement of clinical NLP in healthcare.

Keywords: Corpus (Creation, Annotation, etc.), Document Classification, Text categorisation, Neural language representation models

1. Introduction

In the realm of clinical NLP research, publicly available datasets are a rarity, primarily due to privacy and ethical concerns. The pivotal release of MIMIC-III (Johnson et al., 2016) has been decisive for the reproducibility of research results and progress in the area of clinical NLP. A tremendous amount of influential research was published based on the dataset. To support such research efforts, MIMIC-IV (Johnson et al., 2020) was released, expanding the dataset to include patient data up to 2019. This release increased the number of unique patients by approximately six times, providing a richer and more extensive dataset. Recently, van Aken et al. (2021) introduced the task of predicting clinical outcome from clinical text, only incorporating information available at admission time. Predicting clinical outcome is essential in modern healthcare. It serves as a preventive tool, aiding doctors during the differential diagnosis process by identifying potential risks and symptoms, as well as assisting hospitals in proactive capacity planning. In this investigation, we present a clinical outcome prediction dataset derived from MIMIC-IV. Additionally, we offer initial observations regarding the efficacy of models trained on MIMIC-III when applied to our dataset, paying particular attention to data drift from the inclusion of more recent documents from MIMIC-IV. In addition to understanding complex linguistic aspects like relationships, ambiguity, negations, abbreviations, and other language intricacies, clinical documentation of diagnoses and procedures presents extra challenges for these models when applied to real-world tasks. To determine the correct diagnosis and procedure codes that are associated with the patients' clinical note, professional coders need to manually assign codes, organized in a standardized taxonomic hierarchy like ICD-9, ICD-10 or ICD-11 (International Statistical Classification of Diseases and Related Health Problems (ICD)) [1]. The ICD system is widely established and is used for billing and clinical documentation purposes. Searle et al. (2020) highlight that clinical text, combined with the often stringent time constraints placed on clinical coders, adds to the likelihood of errors and inconsistencies in this process. Moreover, the ICD-Code taxonomy is updated regularly, which results in inconsistencies between each major revision. The goal of the regular updates is to capture an expanding range of health conditions, procedures, and therapies with increasing granularity. Major updates to the ICD system are not fully backwards compatible, which makes using clinical data across time challenging. We present the first analysis of different aspects of the generalization and robustness capabilities of state-of-the-art models trained on MIMIC-III data and evaluated on MIMIC-IV. Our evaluation focuses on the following aspects:

- The implementation of the major revision of ICD-9 to ICD-10 that is partially present in the MIMIC-IV data

[1] https://www.who.int/standards/classifications/classification-of-diseases

LREC-COLING 2024, pages 4381–4391
20-25 May, 2024.

- The effect of changes in clinical documentation standards or guidelines
- The different hospital wards in which the text is produced, i.e. the emergency department and the ICU.

Contribution We summarize our contribution as follows:

1. Generation of an outcome prediction dataset utilizing MIMIC-IV data [2].
2. Application of techniques to establish correspondence between existing MIMIC-III documents within MIMIC-IV, facilitating comparability for models previously trained on the MIMIC-III outcome prediction task.
3. Rigorous selection of dataset splits designed to facilitate the reuse of the original MIMIC-III test dataset for outcome prediction purposes (van Aken et al., 2021).
4. Evaluation and comparison of existing state-of-the-art models for clinical outcome prediction with regard to data drift.

2. Related Work

MIMIC-III The freely available Medical Information Mart for Intensive Care v1.4 database, also known as *MIMIC-III* (Johnson et al., 2016), fueled medical computational science research since 2016 with thousands of citations and provided a valuable foundation to an abundance of publications. MIMIC-III contains de-identified electronic health record data including textual discharge summaries in English of the Beth Israel Deaconess Medical Center (BIDMC) in Massachusetts ranging from 2001 to 2012.

MIMIC-IV The *MIMIC-IV* (Johnson et al., 2020) dataset is the successor to MIMIC-III and contains MIMIC-III data from 2008 to 2012 as well as new data collected from 2013 to 2019. Furthermore, in addition to ICU data, it also contains data from the BIDMC emergency department. The incorporation of emergency department data significantly enhances the phenotypic diversity within MIMIC-IV, capturing a broader spectrum of patient profiles, including those not requiring critical care interventions.

Coding Differences Between MIMIC-III and MIMIC-IV MIMIC-III uses the ICD-9 coding standard to encode diagnoses and procedures. MIMIC-IV relies on the newer ICD-10 standard for data collected between 2013 and 2019, as well as ICD-9 for data before 2013. While ICD-9 contains 3,878 procedure as well as 14,567 diagnosis codes, the much more fine-grained ICD-10 contains 71,920 procedure and 69,832 diagnosis codes. Researchers and practitioners who seek to evaluate tasks that rely on the medical coding information, like clinical outcome prediction, have to tackle issues that come with multiple coding standards, e.g. ambiguous label spaces. One approach is to perform independent evaluations on each respective label space, e.g. Bornet et al. (2023) use only the documents annotated with ICD-10 codes. However, depending on the type of code (diagnostic or procedures) MIMIC-IV contains up to 66.7% of ICD-9 coded data which forms the majority of the dataset (s. Section 3 for more details). To enable comparability of models on both coding standards present in MIMIC-IV, they need to be combined into a common label space. For ICD-10 as well as ICD-9, there are sophisticated mapping mechanisms provided by the Centers for Medicare and Medicaid Services as well as the Centers for Disease Control and Prevention called *"General Equivalency Mappings"* (GEMs) [3].

Clinical Outcome Prediction Task van Aken et al. (2021) proposed a task to predict a patients' outcome from a clinical note written at admission time. They proposed a dataset based on MIMIC-III with four tasks: 1. Diagnosis Prediction, 2. Procedure Prediction, 3. Length-of-Stay and 4. In-Hospital Mortality Prediction. With the help of trained medical professionals, they chose to extract a list of sections from the discharge summaries from MIMIC-III that are known at admission time. The resulting documents, called *Admission Notes*, only contain information that is likely to represent the knowledge about a patient's state at hospital admission. Thus, they can be used to evaluate the task of clinical outcome prediction. This differentiates the task of clinical outcome prediction from the task of ICD coding (Edin et al., 2023; Mullenbach et al., 2018), where the full document is used.

Clinical Outcome Prediction Approaches To solve the outcome prediction task, several approaches have been proposed, such as the *CORe* (Clinical Outcome Representations) model (van Aken et al., 2021) that aims to learn clinical outcome representations of admission notes by continuing pre-training on discharge summaries. Besides the continued work of the authors, like using prototypical networks for more interpretable predictions (van Aken et al., 2022) or behavioral testing

[2] https://github.com/DATEXIS/MIMIC-DataDrift

[3] https://www.cms.gov/medicare/coding-billing/icd-10-codes/2018-icd-10-cm-gem

PRESENT ILLNESS: 58yo woman w/ hx of hypertension, AFib on coumadin presented to ED with headache and chest pain. Husband reports states that patient has been complaining of headache for 1 days, chest pain for 3 days and has lost consciousness 2 days ago for a minute
MEDICATION ON ADMISSION: The Preadmission Medication list is accurate and complete. 1. Aspirin 120 mg PO DAILY, 2. Simvastatin 20 mg PO QPM
PHYSICAL EXAM: Vitals: P: 82 R: 12 BP: 140/75 SaO2: 95%
Cardiac: RRR
atraumatic, normocephalic Pupils: 4-3mm. Abd: Soft, BS+ Extrem:
Warm and well-perfused.
FAMILY HISTORY: non-contributary
SOCIAL HISTORY: Lives together with husband

Figure 1: Example of an admission note from MIMIC-IV.

(van Aken et al., 2022), there have been many works on improving on the task. Winter et al. (2022) enhance encoder models with external information from knowledge graphs by injecting it into redundant attention heads. Naik et al. (2022) augment the classification process by adding retrieved documents from PubMed. Grundmann et al. (2022) augment encoder models to use optional previously known diagnosis codes to support the prediction. Taylor et al. (2023) employ a prompt learning approach. Ji and Marttinen (2021) use task-specific embeddings. Papaioannou et al. (2022) enhance the encoders by applying cross-lingual knowledge transfer. Also, the outcome prediction dataset proposed by van Aken et al. (2021) was used for numerous other tasks besides clinical outcome prediction, e.g. evaluation of trustworthiness of synthetic data (Belgodere et al., 2023) or analyzing the interpretability of classifiers (Naylor et al., 2021). To the best of our knowledge, the approach of van Aken et al. (2022), based on prototypical networks, provides the best performance on the diagnosis prediction task.

3. Preparing the Dataset

MIMIC-III consists of in total 53,423 patient stays and 38,597 unique patients, resulting in an average of 1.38 stays per patient. It covers a time span from 2001 until 2012. The *MIMIC-IV* (Johnson et al., 2020) dataset can be understood as the successor to MIMIC-III and contains MIMIC-III data from 2008 to 2012 as well as new data collected from 2013 to 2019. Furthermore, in addition to ICU data (66,239 stays with 50,920 unique patients), it also contains data from the BIDMC emergency department (431,231 stays with 180,733 unique patients).

3.1. Extraction of Admission Notes

Following van Aken et al. (2021) we adapt the extraction methodology and extract the following sections from the discharge summaries in MIMIC-IV: *Chief Complaint*, *Present Illness*, *Medical History*, *Admission Medications*, *Allergies*, *Physical exam*, *Family History* and *Social History*. In contrast to MIMIC-III, the section *Social History* is often not present in MIMIC-IV. We provide an example of an admission note, modified for anonymity reasons in Figure 1.

3.2. Matching

MIMIC-IV contains parts of MIMIC-III due to the overlap during the years 2008-2012. To establish comparability with the outcome prediction dataset by van Aken et al. (2021), it is necessary to exclude any document from the MIMIC-III training dataset split in our MIMIC-IV test dataset splits. Therefore, we need to identify the respective documents from MIMIC-III in MIMIC-IV. However, identifying the content of MIMIC-III in MIMIC-IV is not trivial. First, all the unique identifiers, e.g. of patients and admissions, have been re-generated so that no direct match is possible, and it is unknown whether the entire ICU subset from that timeframe is part of MIMIC-III or not. Second, MIMIC-IV changed the de-identification process. Instead of replacing HIPAA (Health Insurance Portability and Accountability Act) defined data by random identifiers, all discriminating data is replaced with three underscores. Furthermore, MIMIC-IV only contains information about the year of a patients' admission instead of the day, week and season identifiers present in MIMIC-III. Patients are now grouped by a so-called anchor-year-group (e.g. 2011-2013) that indicates when the patients' first admission took place. MIMIC-IV contains four anchor year groups in total: 2008-2010, 2011-2013, 2014-2016 and 2017-2019. Because the data collection for MIMIC-III stopped in 2012, there cannot be any MIMIC-III admission in the anchor-year-groups after 2013. In consequence, we focus on the groups 2008-2010 and 2011-2013 which must contain admissions from MIMIC-III. However, this data also includes all patients and admissions from 2013, not present in MIMIC-III. Furthermore, Johnson et al. (2022) provide a dataset split called *"MIMIC-III Clinical Database CareVue subset"* based on MIMIC-III that is guaranteed to be not part of MIMIC-IV and mostly consists of documents from before 2008. It uses the same unique identifiers as MIMIC-III. By removing every document from MIMIC-III that is also part of CareVue, we can reduce the number of false positives resulting from our matching approaches. The resulting difference contains 23,294 patients and 32,140 admission notes. We use the following two matching approaches to identify those remaining MIMIC-III admission notes in MIMIC-IV:

Matching by Earliest Possible Year Patients are assigned an anchor-year-group according to their first admission. In addition, a patient receives an anchor year with added noise. Each admission of a patient has a discharge time feature that is based upon the anchor year. First, we identify patients in the first two anchor-year groups (2008-2010 and 2011-2013) and filter the admissions. We subtract the difference between the anchor year and the minimum of the anchor-year-group from the year of discharge of an admission. This enables us to filter out all admissions from before 2013. We refer to this matching approach in the following as EPY.

Feature-Based Matching by Length-Of-Stay, Diagnoses and Procedures To increase the filtering precision, we match admissions that share the following features: Diagnoses, procedures and length-of-stay. We define the length-of-stay to be the difference between the discharge and admission time of a patient stay. Furthermore, we match by using the respective timestamp of the admission, but only use the hour, minutes and seconds as they are kept the same as in MIMIC-III. Year, month and day features are obscured in MIMIC-IV. We consider an admission to be part of MIMIC-III if all the mentioned features match. In the following, we refer to this matching approach as FBM.

Matching Quality Estimation To guarantee the precision of our chosen matching methods to identify matching admission notes in MIMIC-III and MIMIC-IV, we perform a range of validation experiments. In a quantitative evaluation, we choose random text sequences from the discharge notes of two matching admissions and evaluate whether we can find overlaps between the two. We find that feature-based matching of the admission does not contain false positive matches. Matching by earliest possible year EPY overlaps mostly with the feature-based matching FBM but contains many false positives. Further, we perform a qualitative evaluation by analyzing random examples from matched admission notes. We find that the set $EPY \backslash FBM$ contains different admission- and discharge timestamps for the matched admissions. We consider those admissions a false positive match. $FBM \backslash EPY$ shows us that MIMIC-IV contains around 3,000 patients which are not in the ICU module but part of MIMIC-III. We also find 128 documents that are not matched via the EPY matching. Finally, we find 60 examples in the set of MIMIC-III\MIMIC-III$_{(CareVue \cap EPY \cap FBM)}$. Five examples of this set did not have discharge summaries, and 55 could not be matched. We conclude that the combination of our chosen matching methods provides sufficient precision that we can say that, up to our knowledge, there are no false positive matches resulting from our matching approaches. Using our matching we can identify 99.37% of the remaining subset in MIMIC-IV.

3.3. Dataset Splits

In order to evaluate data drift, we create six different splits out of MIMIC-IV that we use for evaluation, presented in Table 1. $III-test$ is the original test split of van Aken et al. (2021) which serves as a baseline. $IV_{III-test}$ contains all the data from MIMIC-IV that we identified using our matching algorithm to be also present in the $III-test$ dataset split. This split allows us to compare the effects of the new de-identification scheme used in MIMIC-IV. IV_{HOSP} contains all the admissions from the emergency department / hospital module in MIMIC-IV. It excludes all ICU admissions. This split enables us to evaluate whether models are capable to generalize from seen ICU data to another clinical domain. Analoguesly, the emergency department split, IV_{ICU} uses only the ICU data. This dataset also contains parts of the MIMIC-III training data that we use for training. Finally, $IV_{ICU \backslash III}$ consists of all admissions that are not present in MIMIC-III. Therefore, it can be used to measure the data drift from 2001-2012 to the new data in MIMIC-IV from 2013-2019.

3.4. General Equivalency Mappings

Share of ICD-9 and ICD-10 Codes in MIMIC-IV MIMIC-IV contains data from 2008 until 2019. ICD-10 was introduced in 2012. In consequence, we find examples annotated with both ICD-9 and documents annotated with ICD-10 in MIMIC-IV. The dataset contains 58.2% ICD-9 diagnosis codes, and 66.7% ICD-9 procedures codes. The remaining codes are annotated in ICD-10 format. In order to enable comparable results, we establish a common label space. The available GEMs allow us to convert either from ICD-9 to ICD-10 or vice versa. As ICD-9 does not directly map to ICD-10 and ICD-10 is more fine-granular, the GEMs define four different match types: *"NO MAP", "IDENTICAL", "APPROXIMATE" and "COMBINATION"*. *"NO MAP"* and *"IDENTICAL"* represent that there is no mapping between both codes, or that there exists a direct one-to-one mapping. *"COMBINATION"* often can be resolved automatically, as it means that a certain combination of codes lead to either a different combination of codes or a single code. The same applies for *"APPROXIMATE"* which means that often there is a one-to-one mapping that fails to be an identical match due to less specificity in the target system. However, *COMBINATION* and *APPROXIMATE* can also lead to non automatically resolvable mappings that require addi-

Dataset	øDiag./Adm.	øProc./Adm.	øLength-of-Stay	øMortality
$III - train$	11.17	4.16	10.14 days	10.41%
$III - test$	11.27	4.10	10.00 days	10.44%
$IV_{III-test}$	13.24	3.73	8.99 days	9.07%
IV_{HOSP}	10.68	2.17	4.60 days	0.72%
IV_{ICU}	14.73	3.23	3.12 days	9.39%
$IV_{ICU \setminus III}$	15.85	3.05	3.20 days	9.60%
$IV_{ICU_9 \setminus III}$	15.59	2.93	3.29 days	-
$IV_{ICU_{10} \setminus III}$	16.48	3.32	2.98 days	-

Table 1: Overview of the dataset statistics from MIMIC-III and different splits we choose to generate from the MIMIC-IV dataset. This table shows the average number of diagnoses and procedures per admission and the average length of stay in days as well as the average mortality of an admitted patient. III indicates that this split was derived from MIMIC-III, IV that it is based on MIMIC-IV.

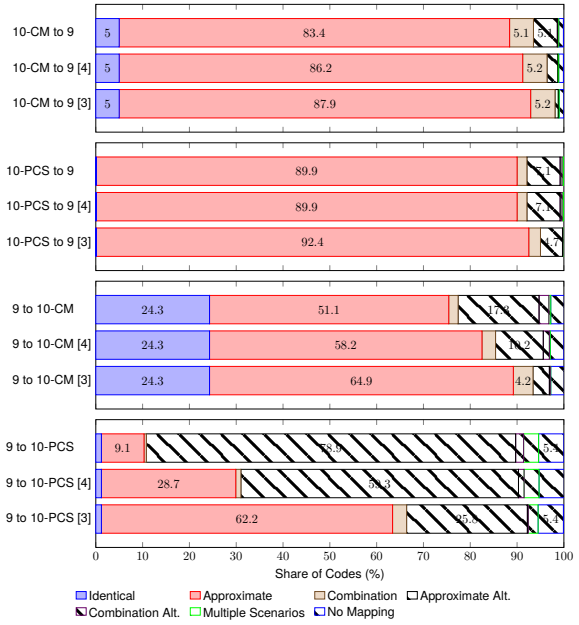

Figure 2: Relative frequencies of each GEM type per GEM table. Automatically non-resolvable types are marked with stripes. The number in square brackets indicates the number of digits per ICD code after grouping. CM refers to diagnosis, PCS to procedure codes.

tional expert knowledge to be resolved. We refer to those non-resolvable types in the following as *"APPROXIMATE ALTERNATIVES"*, *"COMBINATION ALTERNATIVES"* and "MULTIPLE SCENARIOS". *"MULTIPLE SCENARIOS"* refers to one or multiple codes that can be mapped in two or more different ways into the other system.

Application of GEMs - Translation Direction As mentioned in Section 2, the ICD-10 system is much more specific and contains almost five times as many diagnosis codes and 20 times as many procedure codes in comparison to ICD-9. In Figure 2 we show that it is harder to map from ICD-9 to ICD-10, especially regarding procedure codes. We find, that mapping from ICD-9 to ICD-10 leads to more cases of *"APPROXIMATE ALTERNATIVES"*, *"COMBINATION ALTERNATIVES"* and *"MULTIPLE SCENARIOS"* which are not automatically resolvable. This confirms our initial hypothesis that mapping to a more specialized system is harder than the other way around. Mapping from ICD-10 to ICD-9 results in fewer non-resolvable codes, but involves losing some of the specialized and detailed codes from the more recent ICD-10 system. This works both for the procedure codes as well as the diagnosis codes. Therefore, we decide to translate all codes in our experiments from ICD-10 to ICD-9 and evaluate only using the ICD-9 system.

Improving Translation by Code Grouping In Figure 3 we show that 80.5% of the admissions contain at least one diagnosis code and 80.2% at least one procedure code that is not mappable to the respective other system. The ICD system is built hierarchically. This allows us to reduce the complexity of each code following the approach of van Aken et al. (2021). We group the ICD codes by using only their first three digits. In consequence, this increases the number of resolvable codes. This also leads to a denser label space, and the remaining three digits of each code still contain meaningful and helpful clinical information. In Figure 2, we illustrate the impact of grouping to four and three digits, denoted by [4] and [3]. We find that reducing the precision to three digits leads to a large decrease of non-resolvable codes.

Investigation of Remaining Non-Resolvable Codes We further investigate which codes still remain non-resolvable after code grouping. We observe that the occurrence of not resolvable

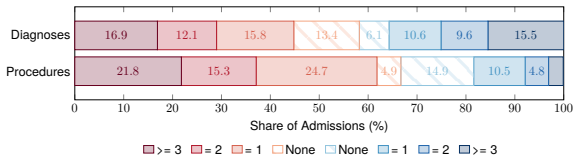

Figure 3: Relative frequencies of MIMIC-IV admissions with non-resolvable codes. 81.2 % contain at least one non-resolvable code in their diagnosis, 74% in their annotated procedures. *ICD-9 in red, ICD-10 in blue*

codes follows a power law distribution. The top five most occurring not resolvable diagnosis codes are responsible for 62.14% of not resolvable cases. An example for this is the ICD-10 code **F32.9** *Major depressive affective disorder, single episode, unspecified*, that can either be translated to **296.20** *Major depressive affective disorder, single episode, unspecified* or **311** *Depressive disorder, not elsewhere classified*. We find that almost all of the identified codes are part of the *APPROXIMATE ALTERNATIVES* category in the GEM mapping.

4. Experiments

We evaluate the effect of data drift on models trained on MIMIC-III using the evaluation datasets defined in Subsection 3.3. We follow van Aken et al. (2021) and evaluate state-of-the-art models on the set of clinical outcome prediction tasks. The tasks have been identified from doctors as a meaningful evaluation in the clinical setting. The set of clinical outcome prediction tasks consists of the following tasks:

Diagnosis and Procedures Prediction The task is to predict the resulting diagnoses or procedures at discharge time, given a patients' admission note. We use the grouped first three digits of the annotated codes. This task is a multi-label classification task.

Diagnosis+ and Procedure+ In addition to the procedure and diagnosis code prediction tasks, we add the four-digit codes and a bag of words representation of each codes' name to the label set. This enables a more precise prediction, as the model can predict either the three digit, four-digit or the bag of words variant of a code. We call this task setting in the following section ICD+, analogue to the implementation of van Aken et al. (2021).

Length-Of-Stay The task is to predict the duration of a patients' stay, given his admission note. To ease the usage of classification models, the length-of-stay is grouped into four categories: less than 3 days, 3 to 7 days, 7 to 14 days and more than 14 days. These four groups were recommended by medical professionals.

In-Hospital Mortality Prediction The task is to predict whether a patient deceases during his hospital stay, given his admission note.

4.1. Models

We evaluate the following models on the task, as they have shown good performance on the outcome prediction tasks on MIMIC-III data and because the models and training code are publicly available. We use the original CORe (Clinical Outcome Representations) model (van Aken et al., 2021), PubMedBERT (Gu et al., 2021) and the ProtoPatient (van Aken et al., 2022) model.

CORe The CORe model [4] is a BioBERT-based model (Lee et al., 2020) that was further pre-trained on admission notes from MIMIC-III and clinical cases from PubMed using a masked-language modeling task. In addition, it was trained on a task similar to next-sentence prediction, where the model has to predict whether two text sequences are part of the same patient note.

PubMedBERT For comparison, we evaluate our tasks on PubMedBERT (Gu et al., 2021), which was pre-trained on 14 million abstracts from PubMed. Unlike other models like BioBERT (Lee et al., 2020), Gu et al. (2021) also pre-trained the tokenizer, thereby enhancing the representation for domain-specific words by preserving them from being broken down into single word pieces. Hence, PubMedBERT generally exhibits superior performance compared to BioBERT. We opted against utilizing Clinical BERT (Alsentzer et al., 2019) due to its pre-training on text sourced from MIMIC-III, a factor that could potentially confer the model with an unjustified advantage.

ProtoPatient In addition to CORe, van Aken et al. (2022) propose a neural network architecture based on prototypical part networks (Chen et al., 2019) and Transformer-based language models. The ProtoPatient model uses label-wise attention to learn one prototype- and attention vector for each diagnosis. A patient's note is mapped to multiple prototype vectors. The combination of attention and prototype vectors enhances the interpretability of the classification. Each prototype vector linked to a diagnosis label highlights a short,

[4] https://huggingface.co/bvanaken/CORe-clinical-outcome-biobert-v1

coherent text sequence from the admission note. ProtoPatient shows state-of-the-art performance on the diagnosis classification task.

4.2. Experimental Setup

We fine-tune all models on the original MIMIC-III training split from van Aken et al. (2021) on the original ICD-9 labels. For all evaluation splits, we remove all admission notes where one or more ICD codes are not automatically resolvable using the code matching described in Section 3.4 and evaluate only on ICD-9 codes. For the CORe model and PubMedBERT we perform a hyperparameter optimization on the same of dataset splits defined by van Aken et al. (2021). We tune the learning rate from $5e-6$ to $1e-4$, warm up steps from 30 to 5000, gradient accumulation steps from 1 to 20, dropout from 0.1 to 0.3 and optimized for maximization of AUROC. For ProtoPatient, we used the hyperparameters proposed by the authors, as they led to reproducible scores from the paper. The ProtoPatient model enhances interpretability on especially many-label classification tasks because it performs the classification on a per-token basis. Therefore, we decided against the evaluation of this model on the mortality and length-of-stay prediction tasks. Furthermore, we only perform more detailed experiments using the CORe model, as it showed similar performance to PubMedBERT. For CORe and ProtoPatient we use the original implementations by van Aken et al. (2021, 2022).

5. Results

Diagnosis and Procedure Task Results Table 2 shows the performance of the chosen models, trained on MIMIC-III, on the respective split of MIMIC-IV and MIMIC-III. A lower relative performance compared to the MIMIC-III test split can be attributed to data drift. We observe that as expected the models do not generalize well from ICU training data to the provided emergency department data in IV_{HOSP}. This applies to procedures as well as the diagnosis code prediction task. For the ICD+ variants, we observe a larger relative performance in comparison to their 3-digit counterparts for all models. PubMedBERT benefits from the added information on the ICD+ procedures task. CORe and PubMedBERT both show reduced performance on the diagnoses ICD+ task. We observe that PubMedBERT performs slightly worse than the CORe model but follows similar performance trends regarding the performance on the MIMIC-IV splits.

Performance Impact on ProtoPatient We find that especially the ProtoPatient model does not generalize well on the MIMIC-IV diagnosis task in comparison to the CORe model. This applies to all subsets and results in an average of -8.62 p.p. (percentage points) of AUROC in comparison to MIMIC-III. From the scores we cannot account this problem to data drift alone as the performance is also worse on the $IV_{III-test}$ split from MIMIC-IV. This means that the new de-identification scheme that has been applied on the MIMIC-IV data is partially responsible for the worse performance of the ProtoPatient model. We hypothesize that the model learns to focus more on very specific keywords and thus overfits on the de-identification schema from MIMIC-III. However, in contrast to the diagnosis task, we observe that the ProtoPatient model still performs well on the procedure code prediction task. Note, that $IV_{III-test}$ splits for the diagnosis and procedure tasks do not fully contain all admission notes from the original III test split due to our matching approach, neither any note from before 2008.

5.1. Impact of GEMs

To assess the effect of using GEMs for translating ICD-10 annotated patient notes to ICD-9, we compare the performance on notes originally coded with ICD-9 and those annotated solely with ICD-10 codes. From the resulting scores in Table 3, we deduce that applying the GEMs has a negative effect on the prediction performance of ProtoPatient on the procedure predictions. CORe, on the other hand, seems to handle the introduced data drift from newer coding standards better. It remains to note that the absolute performance numbers are not comparable due to different label spaces and number of examples in each split. For the diagnosis code prediction, we notice almost no difference in prediction performance.

5.2. Mortality Prediction - Data Drift on Emergency Department Data

We observe in Table 2 that the CORe model, trained on MIMIC-III at first sight, performs well on the mortality prediction task and also seems to generalize on emergency department data (IV_{HOSP}). However, on closer inspection, using Precision, Recall and F1 we observe in Table 4 that the model fails to predict the minority class and only performs well on the majority class. As ICU patients are 5.5 times more likely to decease during their stay, the model favors the prediction of a patients' death, resulting in a low precision score.

5.3. Length-of-stay

Applying MIMIC-III trained models on the length-of-stay task on MIMIC-IV data, we observe that

Model	Task	III	$IV_{III-test}$	IV_{HOSP}	IV_{ICU}	$IV_{ICU \setminus III}$
ProtoPatient	ProtoPatient$_{diag}$	86.09	79.43$_{-6.66}$	76.30$_{-9.79}$	77.95$_{-8.15}$	76.21$_{-9.88}$
	ProtoPatient$_{proc}$	87.97	88.99$_{+1.02}$	86.15$_{-1.82}$	88.29$_{+0.32}$	87.19$_{-0.78}$
CORe	Diagnoses	83.05	81.06$_{-1.99}$	79.38$_{-3.67}$	83.14$_{+0.09}$	79.32$_{-3.73}$
	Diagnoses ICD+	83.21	81.20$_{-2.01}$	78.26$_{-4.95}$	81.69$_{-1.51}$	78.48$_{-4.73}$
	Procedures	87.68	88.75$_{+1.07}$	85.38$_{-2.30}$	88.40$_{+0.71}$	87.63$_{-0.05}$
	Procedures ICD+	88.08	88.66$_{+0.58}$	85.25$_{-2.83}$	88.90$_{+0.82}$	87.19$_{-0.89}$
	Length-of-Stay	72.10	71.69$_{-0.41}$	59.57$_{-12.53}$	71.38$_{-0.72}$	69.69$_{-2.42}$
	In-Hospital Mortality	83.10	82.84$_{-0.26}$	84.83$_{+1.73}$	84.29$_{+1.19}$	82.81$_{-0.29}$
PubMedBERT	Diagnoses	83.61	81.56$_{-2.05}$	79.80$_{-3.81}$	80.76$_{-2.85}$	82.61$_{-1.00}$
	Diagnoses ICD+	80.96	79.80$_{-1.16}$	77.67$_{-3.29}$	78.81$_{-2.15}$	80.98$_{+0.02}$
	Procedures	83.68	85.07$_{+1.39}$	83.13$_{-0.55}$	84.63$_{+0.95}$	86.39$_{+2.71}$
	Procedures ICD+	87.90	88.92$_{+1.02}$	85.62$_{-2.28}$	86.93$_{-0.97}$	88.70$_{-0.80}$
	Length-of-Stay	70.25	70.98$_{+0.73}$	58.98$_{-11.27}$	70.18$_{-0.07}$	69.16$_{-1.09}$
	In-Hospital Mortality	82.45	82.37$_{-0.08}$	84.67$_{+2.22}$	86.21$_{+3.76}$	84.21$_{+1.76}$

Table 2: Evaluation results in macro averaged AUROC. We also present the performance difference compared to the performance on the original III test dataset. ProtoPatient does not generalize to MIMIC-IV data for diagnosis code prediction but manages to generalize well for procedure code prediction.

Model	Task	$IV_{ICU_9 \setminus III}$	$IV_{ICU_{10 \to 9} \setminus III}$
ProtoPatient	Diag.	77.07	77.01
	Proc.	89.62	85.88
CORe	Diag.	79.38	79.74
	Diag+	78.26	78.88
	Proc.	85.38	87.08
	Proc+	85.25	86.60

Table 3: Comparison between the performance on the subset of admissions that use ICD-9 codes and the subset of admissions that use ICD-10 codes. For the ICD-10 subset, the codes are translated to ICD-9. Diagnosis prediction remains unaffected by the change of the coding system. Procedure code prediction performs worse for the CORe model on original ICD-9 codes and slightly worse for ICD-10 codes.

	Alive		**Deceased**		
	Prec.	**Rec.**	**Prec.**	**Rec.**	**F1**
III	93.61	90.87	37.36	**46.73**	66.87
$IV_{ICU \setminus III}$	93.42	95.22	**45.00**	36.84	**67.41**
IV_{HOSP}	**99.49**	**96.80**	6.67	31.55	54.57

Table 4: Recall, Precision and F1 performance of the CORe model on the mortality prediction task. On closer inspection, the model fails to generalize from ICU data to emergency department hospital data.

the performance is similar to the MIMIC-III performance for the ICU data regarding the measured macro AUROC in Table 2. However, the performance drops drastically on the emergency department data (IV_{HOSP}). This behavior is expected because the length-of-stay of a patient in the MIMIC-III training dataset is around twice as long (s. Table 1), thus the training distribution drastically differs from the test data distribution. We also measure the F1 score for each class on the III, $IV_{ICU \setminus III}$ and IV_{HOSP} splits in Table 5. We observe that even though the IV_{HOSP} split follows a different distribution with on average shorter stays than patients in MIMIC-III (III), the models perform surprisingly good at identifying shorter stays. For longer stays, however, the performance decreases drastically, with a trend that especially long stays are very unlikely to be predicted. Furthermore, it is noticeable that the average length-of-stay in the newer MIMIC-IV ICU admissions ($IV_{ICU \setminus III}$) is shorter than in MIMIC-III, which also explains the reduced performance on the ICU test split.

	$F1$			
	$<= 3.$	$> 3 <= 7.$	$> 7 <= 14$	> 14
III	41.98	**46.62**	**38.57**	**37.60**
$IV_{ICU \setminus III}$	43.48	39.79	31.58	33.26
IV_{HOSP}	**53.58**	28.00	19.37	15.90

Table 5: Class-wise F1 of the CORe model on the length-of-stay task. The model does not seem to generalize on the emergency department data.

6. Discussion and Findings

GEMs Negligibly Impact Performance We conclude from our observations in Section 5 that

applying GEMs to automatically map from ICD-10 to ICD-9 does not have a large negative effect on the models' prediction performance in the clinical outcome prediction tasks.

Data Drift Does Not Drastically Affect Diagnosis and Procedure Prediction Performance. In contrast to Yang et al. (2022) we could not observe that our selected models suffer from drastically decreased performance due to data drift with the exception for the ProtoPatient model on the diagnosis classification task. According to Yang et al. (2022), the switch from ICD-9 to ICD-10 and changes in how microbiology samples are taken were the main reasons for performance drop. We suggest that compared to time-series data, it is easier to link clinical features with the text they are associated with in clinical notes. Similarly, the features linked to certain diagnosis codes, even with changes in coding standards over time, tend to stay more consistent. We hypothesize that the new de-identification scheme in MIMIC-IV has a bigger effect on how well the models perform in the evaluations compared to other changes. This indicates that Transformer-based language models trained on MIMIC-III are able to generalize well on MIMIC-IV for the diagnosis and procedure prediction tasks from the outcome prediction benchmark.

Decreased Performance on Emergency Department Data (IV_{HOSP}) As expected, we find that models trained on MIMIC-III ICU data perform significantly worse on the new emergency department data from MIMIC-IV (IV_{HOSP}). Especially for the length-of-stay and mortality prediction tasks, we see decreased performance. Surprisingly, the performance impact on the ICD code classification tasks is lower than expected for CORe and PubMedBERT, with -2.58 p.p. macro AUROC on average.

ProtoPatient Suffers More from Data Drift In contrast to the pure Transformer-based language models CORe and PubMedBERT, we observe a drastically lower performance on the MIMIC-IV diagnosis prediction task for the ProtoPatient model. We find that the de-identification scheme has a large negative impact on the models' performance. This implies that the model hyper-focuses on specific keywords from MIMIC-III, lowering the performance on unseen and structurally slightly different data.

7. Conclusion

In this work, we present the new clinical outcome prediction dataset based on MIMIC-IV that is backwards compatible and thus comparable with MIMIC-III. We also use GEMs to map the label space to ICD-9. We show the effect of data drift on language models trained on outcome prediction tasks like diagnosis-, procedure-, length-of-stay and mortality prediction. We provide empirical evidence that Transformer-based language models trained on MIMIC-III are capable of generalizing to unseen admission notes in MIMIC-IV. Finally, this work should allow researchers to probe the performance of models fine-tunined on MIMIC-III data on MIMIC-IV without any effect of data poisoning in the test split. For future work, we suggest considering to refine the mapping of remaining complex not automatically resolvable ICD codes to enhance the datasets quality and to increase the dataset size. Furthermore, we encourage the work on applying language models on the outcome prediction task, especially interpretable models that can help doctors build an intuition for the models' prediction.

8. Ethical Considerations

Using transformer-based language models to predict patient outcomes from clinical text can be helpful for clinicians and medical professionals. However, it is important to recognize that these text-based models might introduce biases from their training and fine-tuning, thus they shouldn not be blindly trusted. Moreover, the incorporation of billing codes, such as ICD-9 and ICD-10, within datasets like MIMIC-III and MIMIC-IV, inherently introduces biases. Therefore, these codes should be approached with careful consideration as an optimal label space. Of particular concern is the inclusion of both ICD-9 and ICD-10 codes in MIMIC-IV, which amplifies the risk of introducing further biases due to the necessity of mapping between the two systems. This work underscores the persistent challenge of noisy mappings between these systems, which can alter the clinical context and significance of certain annotated codes. It is crucial to note that predicting patient outcomes solely based on clinical text, especially in the absence of supplementary data, poses inherent challenges.

9. Acknowledgements

This work was founded by the German Federal Ministry of Education and Research (BMBF) under grant agreements 01IS23013C (More-with-Less), as well as the grant agreement 01IS23015C (SCM) and the grant agreement 16SV8857 (KIP-SDM).

10. Bibliographical References

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. Association for Computational Linguistics.

Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navartil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. 2023. Auditing and generating synthetic data with controllable trust trade-offs. *arXiv preprint arXiv:2304.10819*.

Alban Bornet, Dimitrios Proios, Anthony Yazdani, Fernando Jaume-Santero, Guy Haller, Edward Choi, and Douglas Teodoro. 2023. Comparing neural language models for medical concept representation and patient trajectory prediction. *medRxiv*, pages 2023–06.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. *arXiv preprint arXiv:2304.10909*.

Paul Grundmann, Tom Oberhauser, Felix Gers, and Alexander Löser. 2022. Attention networks for augmenting clinical text with support sets for diagnosis prediction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4765–4775, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Shaoxiong Ji and Pekka Marttinen. 2021. Patient outcome and zero-shot diagnosis prediction with hypernetwork-guided multitask learning. *arXiv preprint arXiv:2109.03062*.

Alistair Johnson, Tom Pollard, and Roger Mark. 2022. Mimic-iii clinical database carevue subset.

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. *Bioinform.*, 36(4):1234–1240.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453.

Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. 2021. Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff. *arXiv preprint arXiv:2107.05693*.

Jens-Michalis Papaioannou, Paul Grundmann, Betty van Aken, Athanasios Samaras, Ilias Kyparissidis, George Giannakoulas, Felix Gers, and Alexander Loeser. 2022. Cross-lingual knowledge transfer for clinical phenotyping. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 900–909, Marseille, France. European Language Resources Association.

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. Association for Computational Linguistics.

Niall Taylor, Yi Zhang, Dan W Joyce, Ziming Gao, Andrey Kormilitzin, and Alejo Nevado-Holgado. 2023. Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using

self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Betty van Aken, Jens-Michalis Papaioannou, Marcel Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix Gers, and Alexander Loeser. 2022. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 172–184, Online only. Association for Computational Linguistics.

Benjamin Winter, Alexei Figueroa Rosero, Alexander Löser, Felix Alexander Gers, and Amy Siu. 2022. KIMERA: Injecting domain knowledge into vacant transformer heads. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 363–373, Marseille, France. European Language Resources Association.

Janice Yang, Ludvig Karstens, Casey Ross, and Adam Yala. 2022. AI gone astray: Technical supplement. *CoRR*, abs/2203.16452.

11. Language Resource References

Johnson, Alistair and Bulgarelli, Lucas and Pollard, Tom and Horng, Steven and Celi, Leo Anthony and Mark, Roger. 2020. *MIMIC-IV*. [link].

Johnson, Alistair EW and Pollard, Tom J and Shen, Lu and Lehman, Li-wei H and Feng, Mengling and Ghassemi, Mohammad and Moody, Benjamin and Szolovits, Peter and Anthony Celi, Leo and Mark, Roger G. 2016. *MIMIC-III, a freely accessible critical care database*. Nature Publishing Group. [link].

van Aken, Betty and Papaioannou, Jens-Michalis and Mayrdorfer, Manuel and Budde, Klemens and Gers, Felix and Loeser, Alexander. 2021. *Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration*. Association for Computational Linguistics. [link].