Revisiting Clinical Outcome Prediction for MIMIC-IV

Tom Röhr*, Alexei Figueroa*, Jens-Michalis Papaioannou*◇,
Conor Fallon*, Keno Bressem†, Wolfgang Nejdl◇, Alexander Löser*
*DATEXIS, Berliner Hochschule für Technik
†Department of Radiology and Nuclear Medicine, German Heart Center Munich
◇L3S, Leibniz University Hannover
{troehr, afigueroa, michalis.papaioannou, cfallon, aloeser}@bht-berlin.de
bressem@dhm.mhn.de
nejdl@L3S.de

Abstract

Clinical Decision Support Systems assist medical professionals in providing optimal care for patients. A prominent data source used for creating tasks for such systems is the *Medical Information Mart for Intensive Care* (MIMIC). MIMIC contains electronic health records (EHR) gathered in a tertiary hospital in the United States. The majority of past work is based on the third version of MIMIC, although the fourth is the most recent version. This new version, not only introduces more data into MIMIC, but also increases the variety of patients. While MIMIC-III is limited to intensive care units, MIMIC-IV also offers EHRs from the emergency department. In this work, we investigate how to adapt previous work to update clinical outcome prediction for MIMIC-IV. We revisit several established tasks, including prediction of diagnoses, procedures, length-of-stay, and also introduce a novel task: *patient routing prediction*. Furthermore, we quantitatively and qualitatively evaluate all tasks on several bio-medical transformer encoder models. Finally, we provide narratives for future research directions in the clinical outcome prediction domain. We make our source code publicly available to reproduce our experiments, data, and tasks.

1 Introduction

Estimating the future clinical state of a patient upon admission to a medical care facility is a task of critical importance. Clinicians must be able to promptly gauge not only the main affliction of patients, but also all the resources needed to streamline their care. A Clinical Decision Support System (CDSS) aids clinicians in a multifaceted way; for instance, they can interact with a clinician in a conversational manner or they can assist in the diagnosis process by offering discrete suggestions. Generative medical assistants, like AMIE (McDuff et al., 2023), enable clinicians to derive diagnostics and treatments by engaging in a conversation with the language model. One way of communicating these findings is to use the International Classification of Diseases (ICD) taxonomy which is also used by medical practitioners to document the admission of a patient, their stay, and release from a medical care facility. While conversational CDSS can provide reasonable answers and may identify important treatment strategies, their suggestions veer substantially from expert suggestions (Benary et al., 2023). Furthermore, validating these suggestions is difficult, given the arbitrarily large output space of decoder-based transformer architectures such as AMIE. However, it is essential for clinicians to validate the predictions of such systems in order to safeguard the well-being of their patients. Given the discrete space of the ICD taxonomy and the necessity of validation, we argue that classification with encoder models is relevant for the clinical outcome prediction domain.

Clinical Outcome Prediction from Admission Notes. We revisit the clinical outcome prediction (COP) tasks as defined in van Aken et al. (2021). These tasks are all based on the third version of the *Medical Information Mart for Intensive Care* (MIMIC-III)(Johnson et al., 2016). Therefore, in this work, we refer to these tasks as *COP-III*. Since the publication of *COP-III*, a new version of MIMIC has been released, MIMIC-IV (Johnson et al., 2023). MIMIC-IV supersedes the third version with more patient data from the intensive care units (ICU). Additionally, it includes data from patients admitted to the emergency department (ED). This increase in available data, both in quantity and diversity, renders the tasks of *COP-III* obsolete. We present *COP-IV*, an updated and extended set of 6 clinical outcome prediction tasks based on MIMIC-IV. This includes 3 out of 4 *COP-III* tasks adapted for the MIMIC-IV ICU

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 208–217
June 21, 2024 ©2024 Association for Computational Linguistics

and ED splits respectively, as well as a novel *patient routing* task. The patient routing task utilizes the exclusive routing information of MIMIC-IV to predict the first transfer of a patient upon admission. We update the three *COP-III* tasks by adapting the data-processing methods to suit MIMIC-IV. Alongside updating the admission note data, we update the target space from ICD-9 to ICD-10. This provides more relevance for clinicians since ICD-10 is the coding version in use since 2015. We evaluate all *COP-IV* tasks against a selection of open[1] clinical transformer encoder models. Moreover, we compare our results for *COP-IV* and the results of van Aken et al. (2021) for *COP-III* to assess whether the performance for clinical outcome prediction improves with the new data.

Contributions. We summarize our contributions as follows:

- We create novel datasets for several outcome prediction tasks, derived from data in both the intensive care unit (ICU) and the emergency department (ED).
- We introduce a novel *patient routing* task, derived from the patient routing information available in the emergency department module of MIMIC-IV. Resulting in 6 tasks overall, with 3 tasks belonging to ICU and ED prediction respectively.
- We benchmark multiple biomedical transformer encoder models on *COP-IV* and present our qualitative and quantitative analysis.
- We present challenges of *COP-IV* and propose future work directions for clinical outcome prediction.
- We release our source code to reproduce our experiments and datasets[2].

2 Related Work

Bio-medical encoders. In the context of transfer learning, several works explore adapting encoder transformer networks such as BERT (Devlin et al., 2018) into specialized settings.

BioBERT(Lee et al., 2019) presents improved performance in bio-medical text mining tasks, by continuing pre-training a BERT model on full-text and abstracts of research articles from PubMed.

Both *ClinicalBert* and *DischargeBERT* (Alsentzer et al., 2019) further pre-train BioBERT models on full-text notes and discharge notes respectively from the MIMIC-III dataset.

CORe (van Aken et al., 2021) reformulates BERT's unsupervised *next-sentence-prediction* pre-training objective as an *admission-discharge-relation*, tasking a BioBERT model to classify whether a sequence coming from an admission-note relates to the discharge section of the same patient.

In contrast to improving a pre-trained BERT or BioBERT model, *PubmedBERT*(Gu et al., 2020) achieves state-of-the-art results on the majority of bio-medical tasks. This encoder is pre-trained from scratch with a domain-specific tokenizer on a corpus based on PubMed.

Advancements in COP. Naik et al. (2021) augments a PubmedBERT model with document retrieval from a PubMed knowledge base. Grundmann et al. (2022) and Winter et al. (2022) incorporate additional modalities in the form of support sets of ICD codes from prior admissions, and knowledge graph completion tasks respectively. Papaioannou et al. (2022) present knowledge transfer strategies to improve performance for low-resource clinical text datasets in different languages. They show that incorporating clinical text written in multiple languages can complement clinical knowledge missing in smaller datasets, especially for non-frequent diagnoses. Deznabi et al. (2021) augment the text modality with time-series data to improve predictions for in-hospital mortality. van Aken et al. (2022) enhances a PubmedBERT encoder with a prototypical network to not only improve prediction results, but also increase the explainability of predictions.

3 COP-IV Tasks

We revisit the task creation process of van Aken et al. (2021) and update it for the MIMIC-IV data.

3.1 MIMIC-IV: Data preparation

Creation of admission notes. The electronic health records (EHR) available in MIMIC are all associated with medical discharge summaries about the visit of a patient to the hospital. We follow the same pre-processing as in (van Aken et al., 2021), adapted to MIMIC-IV. Hence, we

[1] available on https://huggingface.co/

[2] `https://github.com/DATEXIS/ClinicalOutcomePrediction-IV`

	mean	std	mean	std	total notes
	(words/note)	(words/note)	(sent/note)	(sent/note)	
COP-III-ICU	396.3	233.3	32.5	23.1	**48,745**
COP-IV-ICU	495.6	236.7	26.9	16.1	**59,056**
COP-IV-ED	523.9	265.2	28.5	17.5	**269,573**

Table 1: *COP-III* vs *COP-IV* admission notes details. *COP-III* is based on MIMIC-III, while *COP-IV* is based on MIMIC-IV. The amount of available notes in the ICU increases. ED is not available in MIMIC-III.

keep specific sections in the discharge summaries that are known at admission time, such as: *Chief complaint*, *(History of) Present illness*, *Medical history*, *Admission medications*, *Allergies*, *Physical exam*, *Family history*, and *Social history*. An admission note acts as an input for all tasks; in Figure 1 we present an example. Table 1 demonstrates a comparison of the statistics of admission notes in *COP-III* and *COP-IV*. We observe that the resulting ICU data for *COP-IV* contains 21% more admission notes compared to *COP-III*. In sharp contrast, *COP-IV* offers an additional 269,573 admission notes in the novel ED split. We also remark that for *COP-IV* the average length of an admission note increases, while the number of sentences decreases.

Additionally, note that the clearest difference between MIMIC-III and MIMIC-IV in terms of style is the anonymization scheme. MIMIC-III follows HIPAA[3] for anonymization and identifiable entities are replaced with random identifiers and an indication of the previous content. In contrast, MIMIC-IV replaces all identifiable markers with three underscores: "___"(Johnson et al., 2023). We follow van Aken et al. (2021) and do not mask the de-identified tokens and consider them as part of the admission note.

ICD-10 label space. For the diagnoses and procedure prediction tasks in *COP-III*, the labels are ICD-9 codes. Since MIMIC-IV includes admission notes annotated with ICD-10 codes, for these specific tasks in *COP-IV* we choose to predict only for this newer ICD version. We do this only for the diagnoses and procedures prediction tasks since the remaining tasks are independent of the ICD standard.

3.2 Outcome prediction tasks

Patient routing (PR). We introduce a novel task to *COP-IV*. We construct this task by leveraging routing information for patients accessible in MIMIC-IV, which details patient transfers between different units within the hospital. In the patient routing task, we predict the first hospital unit a patient is transferred to upon admission to the emergency department. Note that we only focus on the first transfer of a patient out of the emergency department, since we predict at the time of admission. Furthermore, we consolidate the labels for the patient's routing information that refer to the same class but differ in their naming. For instance, there are several specific hospital section labels related to surgical procedures, which we group together into *surgery*. This process results in a total of 18 classes (Table 2), making this a multi-class classification task.

Patient Routing Prediction		
	Classes	Number of Samples
COP-IV-ED	18	328,589

Table 2: Novel *patient routing* prediction task summary.

Diagnoses prediction (DIA). The diagnoses prediction task in *COP-IV* involves mapping admission notes to the ICD-10 coding standard. Similar to van Aken et al. (2021), we don't capture the full granularity of ICD-10, and limit ourselves to three-digit codes. This significantly reduces label scarcity, but still retains a relevant level of detail since the codes are organized hierarchically (Choi et al., 2017). As we show in Table 3, the label space grows in size significantly compared to the old version *COP-III*. We apply a multi-label stratified sampling approach (Sechidis et al., 2011) to split the dataset into train/val/test. This ensures that all codes appear in the training set at least once. Furthermore, we restrict multiple admissions for a single patient to be present in the same split, to prevent potential data leakage during training. Diagnoses prediction is a multi-label classification task.

Procedures prediction (PRO). The procedures prediction task in *COP-IV* also involves mapping admission notes to ICD-10. In contrast to the diagnoses prediction task, instead of using only the first 3 digits, we use the first 4 digits. This is due to the differences in hierarchy between the diagnoses and procedure codes in ICD-10. Table 4 contains

[3] Health Insurance Portability and Accountability Act

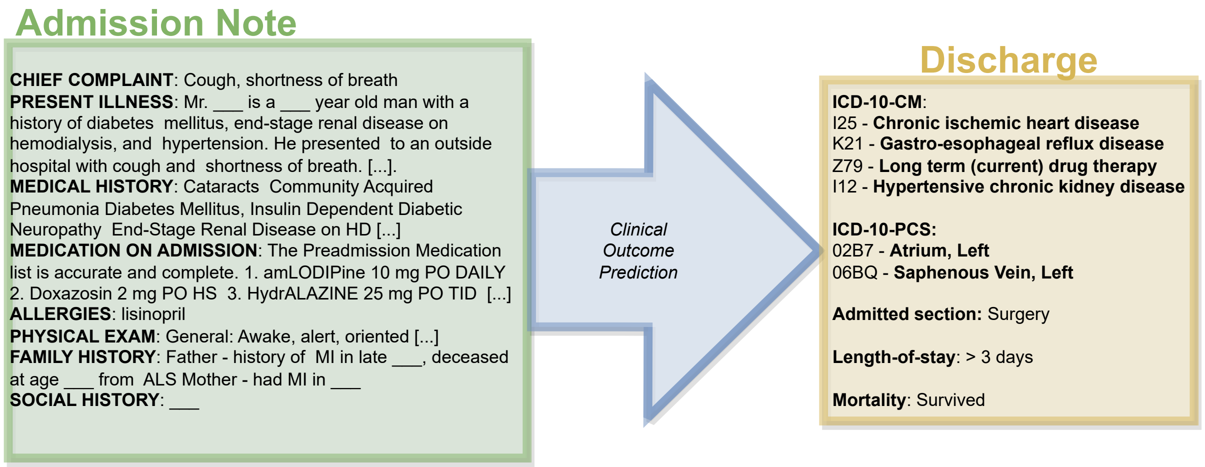
Admission Note
CHIEF COMPLAINT: Cough, shortness of breath
PRESENT ILLNESS: Mr. ___ is a ___ year old man with a history of diabetes mellitus, end-stage renal disease on hemodialysis, and hypertension. He presented to an outside hospital with cough and shortness of breath. [...].
MEDICAL HISTORY: Cataracts Community Acquired Pneumonia Diabetes Mellitus, Insulin Dependent Diabetic Neuropathy End-Stage Renal Disease on HD [...]
MEDICATION ON ADMISSION: The Preadmission Medication list is accurate and complete. 1. amLODIPine 10 mg PO DAILY 2. Doxazosin 2 mg PO HS 3. HydrALAZINE 25 mg PO TID [...]
ALLERGIES: lisinopril
PHYSICAL EXAM: General: Awake, alert, oriented [...]
FAMILY HISTORY: Father - history of MI in late ___, deceased at age ___ from ALS Mother - had MI in ___
SOCIAL HISTORY: ___
Clinical Outcome Prediction
Discharge
ICD-10-CM:
I25 - Chronic ischemic heart disease
K21 - Gastro-esophageal reflux disease
Z79 - Long term (current) drug therapy
I12 - Hypertensive chronic kidney disease
ICD-10-PCS:
02B7 - Atrium, Left
06BQ - Saphenous Vein, Left
Admitted section: Surgery
Length-of-stay: > 3 days
Mortality: Survived

Figure 1: Clinical Outcome Prediction: Given an EHR textual description of a patient admission(left) this task involves determining outcomes (right) such as diagnoses, procedures, hospital section, length of stay, and mortality at discharge.

Diagnoses Outcome Prediction				
	Total	Train	Test	Val
COP-III-ICU	**1,266**	1,201	1,031	906
COP-IV-ICU	**1,447**	1,447	943	943
COP-IV-ED	**1,617**	1,617	1,207	1,198

Table 3: Diagnoses code statistics for *COP-III* vs *COP-IV*. Note that the labels in the *COP-III* diagnoses task are ICD-9 codes and in *COP-IV* these are ICD-10 codes. The label space grows significantly for both splits, ED and ICU.

Procedures Outcome Prediction				
	Total	Train	Test	Val
COP-III-ICU	**711**	672	563	476
COP-IV-ICU	**2,956**	2,956	761	756
COP-IV-ED	**4,137**	4,137	1,242	1,344

Table 4: Procedures code statistics for *COP-III* vs *COP-IV*. The label space grows significantly due to the adoption of ICD-10 in *COP-IV*

a summary of the code distributions for the task. Since in the ICD-10 coding standard there are 19 times more procedure codes than ICD-9[4], the total number of codes increases drastically across the ICU and the ED split. We apply the stratified sampling strategy that we use for the diagnoses outcome prediction task. Procedures prediction is also a multi-label classification task.

Length-of-stay prediction (LOS). Predicting the length of a patient's stay for a visit is beneficial for medical facilities to allocate resources accordingly. As in (van Aken et al., 2021), the length of an ICU stay is defined as the number of days between the admission and discharge of a patient. Unlike van Aken et al. (2021) we focus specifically on the length of a stay of a patient in the ICU, since factors beyond the state of a patient like occupied beds, medical professionals availability, etc. could determine the stay. This information is available in MIMIC-IV and we use the same 4 classes as in *COP-III*: *Under 3 days*, *3 to 7 days*, *1 week to 2 weeks*, and *more than 2 weeks*. We validate these modifications to the task with medical professionals and do not create this task for the ED split. As shown in Table 5, the stay of patients considered in *COP-IV* shifts significantly due to the focus of the stay in the ICU. The majority class is now (*Under 3 days*). Length-of-stay prediction is a multi-class classification task.

Length-of-stay (in days)				
	≤ 3	> 3 & ≤ 7	> 7 & ≤ 14	> 14
COP-III-ICU	5,596	16,134	13,391	8,488
COP-IV-ICU	41,285	11,840	3,986	1,945

Table 5: *Length-of-stay* prediction task for *COP-III* & *COP-IV*. The length of a stay is measured in days. We observe a shift in the class distribution between version III and IV. This task is not applicable to the ED split.

[4] Accessed 28.02.24, https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm

In-hospital mortality prediction. Since a medical professional writes a discharge summary after the visit of a patient, admission sections may contain explicit references to their death. van Aken et al. (2021) applied pattern matching to remove such admission notes. However, in our attempt to replicate this preprocessing method, we found that a rule-based approach to detecting these cases is not reliable. We trained PubMedBERT following this approach; this led to extremely high scores in both AUROC and PR-AUC. Upon closer examination, we still encounter additional patterns (e.g. cessation, passed) that made the decease of a patient explicit. Since we cannot guarantee exhaustive filtering to remove admission notes with such fragments for the MIMIC-IV data, we omit van Aken et al.'s (2021) in-hospital mortality prediction task in *COP-IV*.

4 Experiments

We fine-tune all models in all outcome prediction tasks on both MIMIC-IV splits, except for the *LOS* and *PR* tasks. These tasks are exclusive to the ICU and ED split as mentioned in Section 3. We report performance in *AUROC-macro* as well as in *PR-AUC*. In contrast to van Aken et al. (2021), we include PR-AUC as an additional metric.

While the AUROC provides insight into performance for the majority of the patients, the PR-AUC provides a more balanced view, since it emphasizes the performance of labels that are less frequent in the data.

For comparability, we evaluate all *COP-IV* tasks with the encoder models used in (van Aken et al., 2021), namely BioBERT, CORe, ClinicalBERT, and DischargeBERT. Additionally, we extend this evaluation to PubMedBERT. We conduct a Hyper-Parameter-Optimization (HPO) on PubMedBERT for all tasks for the learning rate and warmup steps using *ray* (Liaw et al., 2018) and (Bergstra et al., 2013). We use the resulting hyperparameters in all experiments. We use early stopping on AUROC with a patience of 5 epochs as in van Aken et al. (2021). We keep a consistent batch size of 50 for all tasks and models. For every experiment, we use a single A100 40GB GPU.

5 Results

We present all experimental results in Table 6.

Overall performance. PubMedBERT outperforms all models across all tasks. BioBERT is the second best performing model, followed by CORe. ClincalBERT and DischargeBERT are the worst performing models.

Domain-specific tokenizer. PubMedBERT is the only model in our work that uses a domain-specific tokenizer. We argue that this is one of the reasons why it is the top-performing model across all tasks. Notably, the average tokenized admission note in MIMIC-IV is longer than 512 tokens. Thus exceding the maximum sequence length for BERT-like models. Therefore, the context window that PubMedBERT processes per admission note contains more information on average compared to the other models.

Pre-training on MIMIC does not bring benefits. PubMedBERT and BioBERT are pre-trained on PubMed. They have not explicitly seen any MIMIC discharge summaries during the pre-training. In contrast, CORe, ClinicalBERT, and DischargeBERT incorporate MIMIC-III data into their training routine, thus exposing the parameters to specific details, writing style, and anonymization scheme. The results suggest that the models do not benefit from pre-training on MIMIC-III. This is highlighted by the fact that BioBERT has a very similar performance. Thus, reinforcing the idea that the domain-specific tokenizer has a much greater impact on the performance of these tasks.

Patient routing. All models achieve high scores for AUROC. In contrast, the results in PR-AUC indicate that all models have difficulties with capturing the hospital units where transfers occur less often. Similar to other tasks, PubMedBert outperforms all other models.

5.1 Performance comparison of CORe on MIMIC-III and MIMIC-IV

To validate that our adaptation of the *COP* tasks to the MIMIC-IV dataset is done correctly, we compare the performance of the CORe model on *COP-III* and *COP-IV*. For *COP-III* we use scores from van Aken et al. (2021) and for *COP-IV* we take the results of the CORe[5] model from our evaluation on the respective task in *COP-IV*. We present this comparison in Table 7. Since the ED split was not available in MIMIC-III, we only compare the

[5] https://huggingface.co/DATEXIS/CORe-clinical-outcome-biobert-v1, accessed 28.02.24

	Task	PR		DIA		PRO		LOS	
Split	Model	AUROC	PR-AUC	AUROC	PR-AUC	AUROC	PR-AUC	AUROC	PR-AUC
ED	BioBERT	93.83	59.33	85.86	14.77	92.87	19.32	-	-
	CORe	93.85	59.55	85.46	14.54	93.57	19.70	-	-
	DischargeBERT	93.87	59.69	84.83	14.29	92.93	19.02	-	-
	ClinicalBERT	93.85	59.19	84.73	14.05	93.18	18.74	-	-
	PubMedBERT	**94.28**	**61.44**	**86.86**	**17.24**	**93.64**	**21.62**	-	-
ICU	BioBERT	-	-	78.71	13.02	86.32	17.44	70.89	36.06
	CORe	-	-	78.06	13.05	85.38	16.10	71.39	36.49
	DischargeBERT	-	-	77.76	12.30	85.25	16.01	70.00	35.70
	ClinicalBERT	-	-	77.02	12.58	84.62	14.86	70.18	35.58
	PubMedBERT	-	-	**79.70**	**15.55**	**87.21**	**18.43**	**71.82**	**36.87**

Table 6: Results of the models for all outcome prediction tasks. Metrics are macro averaged and scores are in %. PubMedBERT is the best performing model for all *COP-IV* tasks. We observe a big gap between AUROC and PR-AUC, signaling the challenges of the long-tail distribution of labels in MIMIC.

	DIA	PRO	LOS
CORe COP-III	83.39	87.15	72.53
CORe COP-IV	78.06	85.38	71.39

Table 7: Comparison of the CORe model's AUROC-macro performance in *COP-III* as reported in (van Aken et al., 2021) and *COP-IV*. The scores are in %. Given the non-existence of the ED split in version III, we compare ICU only. The tasks in *COP-IV* are more challenging, the pre-training on MIMIC-III does not transfer positively to MIMIC-IV.

tasks that relate to ICU data. This also excludes the patient routing task.

Diagnoses and procedures outcome prediction. *COP-III* and *COP-IV* have different label spaces for diagnoses and procedures. We use ICD-10, whereas *COP-III* uses ICD-9. van Aken et al. (2021) reports better performance for both tasks. We argue that this performance gap might be due to the larger code space of ICD-10 compared to ICD-9 (Cartwright, 2013). Additionally, since *COP-IV* uses only ICD-10 codes, we are limited to a fraction of the total amount of summaries available in MIMIC-IV for the ICU split. Roughly 60% of admission notes in this split are annotated with the ICD-9 standard, hence this results in significantly fewer notes for training in *COP-IV* than in *COP-III*.

Length-of-stay. The similar scores for the CORe model in *COP-III* and *COP-IV* in Table 7 indicate that the length-of-stay task is still challenging, despite the modification aimed at focusing on the ICU stay. As previously noted, this leads to a shift of the label distribution, with the majority of patients experiencing shorter stays compared to the *COP-III* task. We argue that this shift in the distribution of the labels could be a factor explaining the lower scores for the task in *COP-IV*. Additional challenges at predicting the length of stay of a patient come from factors such as *employment* or *marital status* which may not be mentioned in a clinical admission note (Khosravizadeh et al., 2016).

6 Discussion & Future Work

6.1 Multi-label outcome prediction

The performance reported in Table 6, shows that the AUROC and especially the PR-AUC metric for the DIA and PRO tasks have a large room for improvement.

Critical long-tail. In Figure 2 we present the label distribution for the complete ED split in MIMIC-IV. It is worth noting that only 100 labels (6% of all labels) are annotated in approximately 67% of the data, whereas the remaining 1,517 labels (94% of all labels) are distributed among the remaining 33% of the samples. We observe the same behavior in the ICU split. We expand the evaluation of PR-AUC of PubMedBERT for class groups depending on their frequency. Figure 3 demonstrates that the model achieves poor PR-AUC performance in the tail of the distribution and improves towards the head. This behavior in PR-AUC emphasizes a weakness of current methods

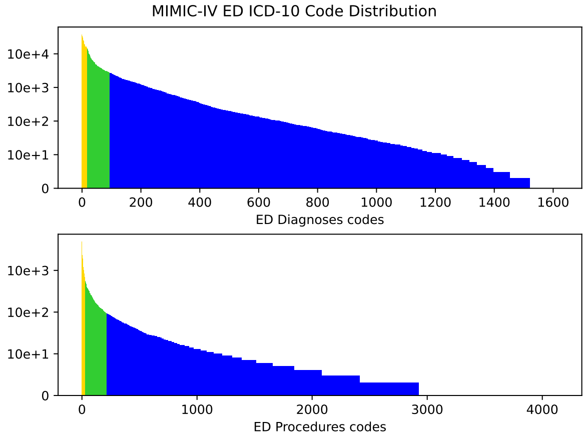

Figure 2: ICD-10 code distribution for the MIMIC-IV ED split. Each one of the 3 colors indicates 33.3% of total samples highlighting a pronounced long tail.

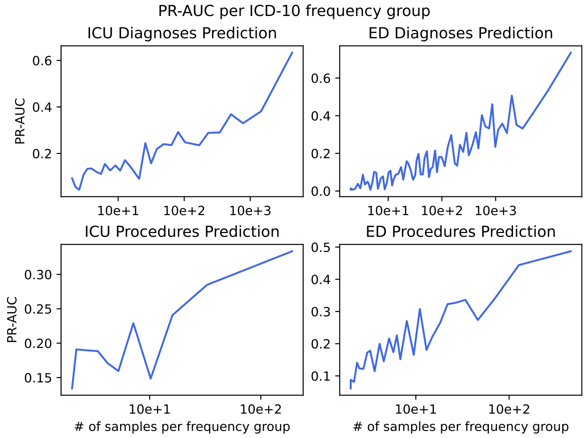

Figure 3: PR-AUC in % measured on groups of labels depending on their frequency in the data. Performance in the long tail is generally poor while it improves greatly for the more frequent labels.

since the majority of the labels reside in the tail.

Label-space. The larger code space in ICD-10 in comparison to ICD-9 further exacerbates the class imbalance present in the multi-label outcome prediction tasks (DIA & PRO).

Annotation Moreover, labels in MIMIC exhibit annotation inconsistencies; in practice the most frequent labels are under-annotated (up to 35%) (Searle et al., 2020). Therefore, some correct predictions made by models will conflict with an incomplete ground truth.

6.2 Qualitative analysis on Patient routing

For the novel patient routing task, we conduct an additional analysis on diversity and identify potential gaps for different populations. Next, we further discuss the difference in performance that we observe in hospital care units. In Figure 4 we disaggregate the PR-AUC for variables such as gender and marital status, as well as admission type and care unit.

Demographic variables. We observe that predictions for male patients are worse by a significant margin. A possible reason could be the additional amount of time spent by women on average for physical exams and patient questions when visiting a doctor (Tabenkin et al., 2004), thus producing more relevant information during the anamnesis. This may result in richer admission notes for women. The *marital status* shows an impact on widowed patients. The average patient is 78 years old, which is 18 years older when compared to the other categories. Given that the age of patients has an impact on other tasks (van Aken et al., 2021; Khosravizadeh et al., 2016), we argue that it has an impact on patient routing as well. For all other classes, the marital status does not seem to influence the outcome.

Admission type PubMedBERT achieves its best performance with admissions that come through physician referrals. Such referrals may contain relevant information to route patients to the corresponding care unit. Walk-ins and Emergency Room (ER) admissions may prioritize immediate care over EHR documentation. Therefore, we argue that in such cases, routing information might be incomplete.

Performance of care units. We observe that performance is not directly coupled to the class distribution. In Figure 4 bottom right, we present the PR-AUC for each care unit, sorting them (from left to right) by the number of occurrences in the data. For instance, *psychiatry* (dark green) and *obstetrics* (dark orange), where PR-AUC is significantly above the average, are units that less present in the data. We argue that for this task performance is determined by the specificity in the admission notes relevant to each care unit and less so by the class frequency. The fact that the *observation* (pink) category is the worst performing reflects the inherent uncertainty of this care unit. We argue that since the symptomatology is not as clear as for other care units (pregnancy in obstetrics), models have more difficulties in routing the patient to the right care unit.

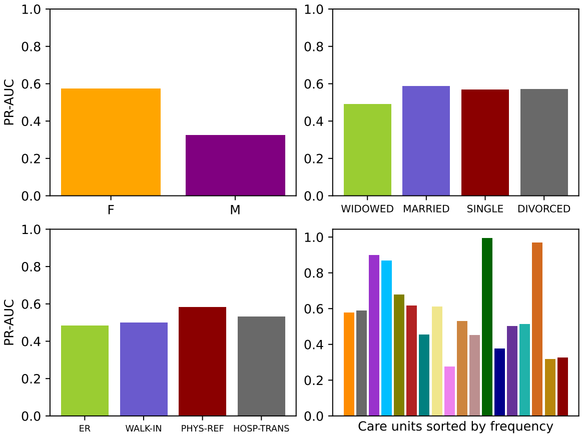

Figure 4: PR-AUC of the patient routing task disaggregated by **Top** demographic variables: Gender and Marital status, **Bottom** Admission type and Care units. A large gap between genders exists. Physician referrals route best. Frequency and marital status of classes are not directly coupled with prediction performance.

6.3 Future Work

Our work aims to be a resource for future research in clinical outcome prediction. We propose future work directions as follows:

ICD code imbalance. We see a very pronounced room for improvement in PR-AUC performance due to the distribution of the labels in the data. We believe that models designed to tackle this premise are needed since it's an inherent feature in the distribution of real-world clinical data. This could be accomplished with novel architectures beyond transformers, or further strategies to integrate complementary knowledge.

Label inconsistency. MIMIC is the best publicly available EHR data and contains annotation deficiencies. We believe that a great effort towards consistent labeling is needed. Potential avenues of data augmentation could come from leveraging generative methods to rephrase and augment existent verified high-quality data.

Evaluation on other datasets. Much of the prior research in clinical NLP has centered around MIMIC. However, evaluating on alternative datasets is crucial. We noticed in our *COP-IV* experiments how models did not benefit from pretraining on MIMIC-III. We believe that these signs of overfitting could be mitigated with broader evaluations using clinical text sourced in different clinics, specialties, and languages.

Multimodal patient representation. Although most modalities relevant to medical practitioners can be expressed in natural language, there are numerous additional modalities available not only in MIMIC but also in other domain datasets. We believe that enriching the textual representations of transformers with multi-modal data could be beneficial for the outcome prediction tasks.

Novel outcome prediction tasks. In practice, outcome prediction consists of a very broad set of possible tasks. Our novel patient routing task is just one example. We expect that additional tasks would provide valuable insights into the strengths and weaknesses of models employed in real-life clinical settings.

7 Conclusion

In this work, we introduce *COP-IV*, a clinical outcome prediction set of tasks based on MIMIC-IV, which updates *COP-III*. In addition, we introduce the novel task of *patient routing* at admission time to clinical outcome prediction. We evaluate qualitatively this task for various patient demographics, as well as hospital care units. We explain in detail our preprocessing approach to reproduce the *COP-IV* tasks. Furthermore, we present a comprehensive evaluation of several bio-medical encoder models and discuss their weaknesses, as well as challenges such as the pronounced class imbalance. Moreover, we give relevant insights into data distribution shifts between *COP-III* and *COP-IV*. Lastly, we propose future research directions for clinical outcome prediction. We release our source code to reproduce the data for our benchmark, experiments, and results.

Acknowledgements

We would like to thank the reviewers for their helpful suggestions and comments. Our work is funded by the German Federal Ministry of Education and Research (BMBF) under the grant agreements 01|S23013C (More-with-Less), 01|S23015A (AI4SCM) and 16SV8857 (KIP-SDM). This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 528483508 - FIP 12, as well as the European Union under the grant project 101079894 (COMFORT - Improving Urologic Cancer Care with Artificial Intelligence Solutions).

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. 2023. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

J Bergstra, D Yamins, and D D Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *TProc. of the 30th International Conference on Machine Learning (ICML 2013*.

Donna J Cartwright. 2013. ICD-9-CM to ICD-10-CM codes: What? why? how? *Adv. Wound Care (New Rochelle)*, 2(10):588–592.

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 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, Online. Association for Computational Linguistics.

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. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, 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(1):160035.

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. *Electron. Physician*, 8(10):3042–3047.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards accurate differential diagnosis with large language models.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2021. Literature-augmented clinical outcome prediction. *CoRR*, abs/2111.08374.

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 Ibrahim, and Richard 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*, pages 76–85, Online. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hava Tabenkin, Meredith A Goodwin, Stephen J Zyzanski, Kurt C Stange, and Jack H Medalie. 2004. Gender differences in time spent during direct observation of doctor-patient encounters. *J. Womens. Health (Larchmt)*, 13(3):341–349.

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 *LREC*, pages 363–373. European Language Resources Association.