Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ICD2Bert: ICD Bert Encodings for Clinical Outcome Prediction on Routine Health Insurance Data

Jonas Harriehausen
University of Giessen
Miriam Cindy Maurer
University of Göttingen
Jacqueline Michelle Metsch
University of Giessen
Zully Ritter
University of Göttingen
Lisa Weller
aQua-Institute
Thorsten Pollmann
aQua-Institute
Matthias Kretzler
BKK Dachverband
Thomas Grobe
aQua-Institute
Anne-Christin Hauschild
anne-christin.hauschild@uni-giessen.de

University of Giessen

Article

Keywords: natural language processing, NLP, machine learning, deep learning, transformer, bert, artificial intelligence, AI, electronic health record, EHR

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10725344/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ICD2Bert: ICD Bert Encodings for Clinical Outcome Prediction on Routine Health Insurance Data

Jonas Harriehausen[1], Miriam Cindy Maurer[2,3], Jacqueline Michelle Metsch[1], Zully Ritter[2,4], Lisa Weller[5], Thorsten Pollmann[5], Matthias Kretzler[6], Thomas Grobe[5], Anne-Christin Hauschild[1,2*]

[1*]Department for Predictive Deep Learning for Medicine and Healthcare, Justus Liebig University Giessen, GießenGermany.
[2]Department of Medical Informatics, University Medical Center, Göttingen, Germany.
[3]Campus Institute Data Science (CIDAS), University of Göttingen, Göttingen, Germany.
[4]Emergency Department, University Medical Center Göttingen, Göttingen, Germany.
[5] aQua Institute for Applied Quality Improvement and Research in Health Care, Göttingen, Germany.
[6] BKK Dachverband, Berlin, Germany.

*Corresponding author(s). E-mail(s): anne-christin.hauschild@uni-giessen.de;
Contributing authors: jonas.harriehausen@uni-giessen.de;

Abstract

While ICD codes are the universal standard for documenting clinical diagnostics and procedures for billing and are therefore widely adopted in clinical practice, due to their specific biases, they remain one of the most challenging data modalities for machine learning. Existing BERT-based models for ICD have introduced domain-specific modifications to better handle this modality; however, most of them have never been compared under controlled conditions in combination with large-scale real-world insurance data from Germany. To address this gap, we propose and benchmark ICD2BERT, a standard BERT model pretrained on ICD

code sequences without domain-specific architectural modifications such as age or visit embeddings. ICD2BERT serves as a controlled baseline for evaluating the benefits of such adaptations. We then systematically compared ICD2BERT with Med-BERT, BEHRT, Pat2Vec, CLMBR, and a categorical baseline across three datasets. The datasets MIMIC-IV, EHRSHOT, and KI-THURST were selected to evaluate model performance across different healthcare systems, data scales, and transfer learning settings. We found that domain-specific modifications to BERT do not improve performance over ICD2BERT, suggesting that pretraining, data quality, heterogeneity, and size matter more than architectural complexity. Remarkably, a non-contextual categorical baseline proved competitive across most tasks, matching BERT performance in several settings. In datasets spanning multiple ICD code standards, older code versions remain critical to model performance, and their removal results in a substantial drop in performance. Transfer learning is most effective when the source dataset encompasses all code types present in the target, but degrades substantially in the reverse direction. Finally, Pat2Vec is largely superseded by BERT except in very low-data settings. Together, these findings suggest that practitioners should carefully weigh the computational cost of transformer-based models against simpler baselines before committing to complex architectures.

Keywords: natural language processing, NLP, machine learning, deep learning, transformer, bert, artificial intelligence, AI, electronic health record, EHR

1 Introduction

Electronic Health Records (EHRs), particularly in the form of International Classification of Diseases (ICD) codes[1], are a standardized, widely adopted[2] way to record patient information. As an international standard, ICD codes have been translated into 43 different languages and adopted in 118 different countries[3]. Although originally designed for billing and administrative purposes, their use has since expanded to encompass clinical decision support and computational analysis. In recent years, there have been many advances in medicine and healthcare using artificial intelligence (AI)[4–6]. AI has proven to be on par with, if not better than, trained experts when it comes to imaging data such as MRIs and X-rays[7–10]. In particular, machine learning can leverage standardized EHR records to uncover complex patterns in a patient's medical history that might be missed by human clinicians[10, 11]. Machine learning in medical care has been shown to improve diagnostic accuracy[12] and to tailor treatment plans to each patient's needs[13]. When it comes to text data such as ICD codes, however, the performance of AI depends strongly on the quality and quantity of the input data[11, 13–16], meaning the data must be clearly defined and uniform to support accurate analysis. However, using ICD codes for these tasks presents significant problems[15–20]. Despite their standardized format, ICD codes are discrete, high-dimensional, and sequential in nature, making it difficult to directly apply many machine learning models without appropriate representation learning.

To address these challenges, we evaluate our models across three datasets. MIMIC-IV[21], a large public US hospital dataset spanning both ICD-9 and ICD-10 diagnosis and procedure codes (See Table 1), serves as our primary pretraining dataset. EHRSHOT[22], a small Stanford cohort with extensive patient histories, evaluates few-shot transfer learning. KI-THRUST[23], a dataset containing German routine insurance claims data, is an order of magnitude larger and more diverse than the hospital-derived datasets, but is restricted to a single year and almost exclusively contains ICD-10 diagnosis codes. This dataset tests cross-national and cross-system generalization. Together, these datasets evaluated the models regarding transferability and adaptability to low-data scenarios.

Alongside the model comparison, we conducted ablation studies on the ICD codes themselves. We omitted code categories or reduced the code specificity to determine which part of the data holds important information and which could be discarded to create a simpler model.

As a baseline, we use multi-hot encoding derived from ICD subcategories, a standard, interpretable choice given its competitive performance relative to its simplicity[22, 24].

Our central architecture is BERT, which has been repeatedly adapted for EHR encoding[25–27], used here without domain-specific modifications such as visit or age embeddings. This isolates whether such additions improve on the base architecture, and preserves compatibility with standard transformer explainability tools. We compare this unmodified ICD2BERT against Med-BERT[26] (visit and serialization embeddings, originally trained on ICD-9/10 diagnoses) and BEHRT[25] (age and segment encodings, originally trained on Read[28]/Caliber[29] codes). We also considered G-BERT but excluded it due to its single-visit input structure and confounding GNN component (see Supplementary Discussion).

We further include Pat2Vec[24], a Doc2Vec[30]-based pre-transformer baseline that lets us quantify the performance gap between the transformer paradigm and its predecessor, and CLMBR[22], a transformer pretrained on the full EHRSHOT feature set (including labs and medications). CLMBR serves as a diagnostic reference rather than a direct competitor: its performance when restricted to ICD codes alone reveals how much the benchmark changes when non-ICD modalities are removed.

Our primary contributions to the field are as follows:

1. A systematic, controlled comparison of six ICD encoding strategies across three diverse but complementary datasets.
2. The first direct comparison of prevalent ICD encoding strategies under identical training conditions.
3. Ablation studies identifying which ICD code categories and specificity levels contribute most to encoding performance.
4. A characterisation of the EHRSHOT benchmark under ICD-only conditions, quantifying the performance impact of excluding non-ICD modalities.
5. Making our code and pretrained model available, to make further research as accessible as possible.

2 Results

We evaluated the six encoders on 21 classification tasks across the three datasets: mortality, rehospitalization, and future disease prediction on MIMIC-IV[21] and KI-THRUST[23] (six tasks), and the 15 EHRSHOT[22] few-shot tasks, summarized into four categories (Operational Outcomes, Anticipating Lab Test Results, Assignment of New Diagnoses, Anticipating Chest X-Ray Findings). The Models were trained on either MIMIC-IV or KI-THRUST and encoded patient histories were classified using Logistic Regression and Random Forest with grid-search hyperparameter optimization, evaluated by AUC-ROC (mean and standard deviation over 10 bootstrapped runs for MIMIC-IV/KI-THRUST, 5 runs for EHRSHOT). Full dataset, training, and evaluation details are given in Methods. The dataset composition, including the distribution of ICD-9/10 diagnosis and procedure codes across datasets, is summarized in Table 1.

		MIMIC	KI-Thrust	EHRSHOT
ICD 10	Diagnosis	1031021 (31.58%)	44020813 (100.0%)	767 (0.0%)
ICD 10	Procedure	104466 (3.2%)	0 (0.0%)	19759 (0.05%)
ICD 9	Diagnosis	1851996 (56.73%)	0 (0.0%)	0 (0.0%)
ICD 9	Procedure	277120 (8.49%)	0 (0.0%)	8468 (0.02%)
	Other	0 (0.0%)	0 (0.0%)	41632643 (99.93%)

Table 1: **Distribution of ICD code types across the three datasets.** Total code counts and percentage of all codes within each dataset for ICD-9 and ICD-10 diagnosis and procedure codes.

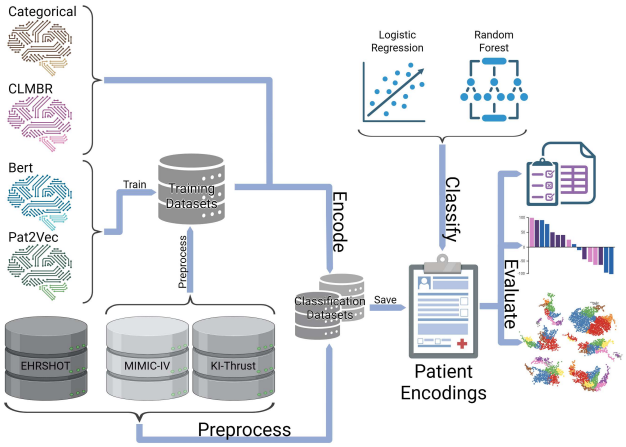


Fig. 1: **Overview of the training and evaluation pipeline.** The left part shows the pretraining and encoding using the different datasets and models. The right part shows the classification and final evaluation.

2.1 Model Architectures

The modes we evaluated on these classification tasks fall into 3 main categories: Transformers, Neural Networks, and Rule-Based (Fig. 1). The BERT model variations and the CLMBR model use the Transformer architecture. The BERT architecture uses bidirectional self-attention, in which each token can attend to the full input context, whereas the CLMBR model uses left-sided attention. BEHRT and Med-BERT, the two BERT variations tested in this study, include input embeddings beyond ICD codes. BEHRT includes the patient's age at the time of recording, while Med-BERT includes the patient's visit count. For ICD2BERT, we use the same architecture as the original BERT paper, adjusting only the encoding dimensionality to match that of the 241-dimensional rule-based model. The other two BERT implementations were also set to this dimensionality for a fair comparison. Finally, we also used Pat2Vec, a Word2Vec-based model architecture. For this model, we used the same hyperparameters as in the original paper and set the encoding dimensionality to 240. We flattened the structured EHR into a quasi-paragraph structure by ordering the EHR Codes by date and severity, then concatenating them into a single sentence. We then concatenated all sentences from the same patient, separated by a [SEP] token. Each patient's paragraph starts with a [CLS] token. The CLMBR model used data in the MEDS schema format, as in the original paper. Using these patient paragraphs, the BERT models encoded the CLS token's last hidden state, yielding higher performance than using the mean of all hidden states. To run all the models, we implemented ICD2BERT model, as well as Med-BERT and BERT, using the Hugging Face library, and Pat2Vec and clmbr were run using the implementations from the original paper. Pat2Vec was implemented in the gensim package and clmbr using the femr and xformer packages.

2.2 Ablation Study

To evaluate the Information contained in each ICD code category, we conducted an ablation study by excluding certain code categories from the MIMIC dataset during training and evaluation. The subsets consisted of only ICD-10 diagnosis codes, only ICD-10 codes, and only diagnosis codes. Furthermore, we evaluated the effect of specificity in ICD codes by limiting the maximum length of diagnosis codes while excluding procedure codes. The length of the ICD codes is limited to either 2 or 3 characters. The final subset in the ablation study consists of ICD codes replaced with their category labels from the categorical encoder, similar to the input data used for the BEHRT model.

The AUC-ROC scores show two clusters of performance scores (Supplementary Table 1). The first cluster consists of the subset containing only ICD-10 codes and the subset containing only ICD-10 diagnosis codes with AUC-ROC scores of roughly 0.44 for future disease, 0.61 for mortality, and 0.53 for rehospitalization prediction. The second cluster contains the remaining subsets with AUC-ROC scores of roughly 0.78 for future disease, 0.8 for mortality, and 0.67 for rehospitalization prediction. The common difference between the two clusters is whether they include the ICD-9 diagnosis codes. As MIMIC-IV contains a small majority of ICD-9 codes compared to ICD 10, the removal of ICD-9 diagnosis codes, in turn removes a majority of the clinical information. On the other hand, removing the procedure codes did not have an impact on performance, which shows that the procedure codes either contain no additional information or no information at all that is relevant for the classification tasks.

These ablated encoders were also evaluated on the KI-THRUST tasks (Supplementary Table 2). Here, the only subsets that showed a decrease in performance were those with limited code length. While the other ablated models maintained performance close to that of the model trained on the full MIMIC-IV dataset.

2.3 Encoder Architecture Comparison

We compared our BERT architecture with other transformer architectures (Med-BERT, BEHRT, and CLMBR), the categorical encoder, and the Pat2Vec architecture. They were all evaluated on the 21 Tasks outlined above.

MIMIC-IV

The BERT architectures trained on the MIMIC dataset have very similar performance, differing in AUC-ROC by at most 0.02 (Table 2). This suggests that the architectural changes in Med-BERT and BEHRT have very little impact on encoding performance, and that the training data has a larger impact on final performance. The Pat2Vec models perform worse than the BERT models, and the pretrained Pat2Vec model performs at a random classification level. The pretrained Pat2Vec model might have underperformed, given its low encoding dimensionality of 10 and the challenge of transferring to a different dataset. The Pat2Vec models trained in this paper do not have these shortcomings. However, they still perform below the BERT models. This suggests that the shallow nature of the Doc2Vec and thereby the Pat2Vec architecture

is not able to capture the information contained in ICD codes as well as transformer models can. The CLMBR model performs slightly better than a random classifier, with AUC-ROC values of 0.56, 0.54, and 0.48 for future disease, rehospitalization, and mortality classification, respectively. This suggests that the model is unable to adapt to the transfer from the Stanford training dataset and the accompanying shift in data modality. The categorical model performs on equal footing with, if not better than, the BERT models, particularly the 640-dimensional variant that additionally includes ICD-9 and procedure codes (see Methods), matching the full code coverage available to the BERT models on MIMIC-IV.

training dataset	model type	dimensionality	future_disease	mortality	rehospitalization
MIMIC-IV	BEHRT	240	0.77 (0.00)	**0.79 (0.00)**	0.66 (0.00)
Ki-THRUST	BERT	240	0.42 (0.01)	0.60 (0.00)	0.51 (0.00)
MIMIC-IV	BERT	240	**0.78 (0.00)**	**0.79 (0.00)**	**0.68 (0.00)**
Ki-THRUST	BERT	768	0.39 (0.00)	0.61 (0.00)	0.51 (0.00)
MIMIC-IV	BERT	768	0.60 (0.00)	0.68 (0.01)	0.52 (0.01)
CLMBR	CLMBR	768	0.56 (0.01)	0.48 (0.00)	0.54 (0.00)
None	Categorical	241	0.40 (0.02)	0.56 (0.00)	0.52 (0.00)
None	Categorical	640	**0.78 (0.00)**	**0.79 (0.00)**	0.66 (0.00)
MIMIC-IV	Med-BERT	240	**0.78 (0.00)**	0.78 (0.00)	0.66 (0.00)
Pat2Vec	Pat2Vec	10	0.51 (0.00)	0.51 (0.01)	0.51 (0.00)
Ki-THRUST	Pat2Vec	240	0.75 (0.00)	0.67 (0.01)	0.57 (0.00)
MIMIC-IV	Pat2Vec	240	0.64 (0.00)	0.72 (0.01)	0.60 (0.00)

Table 2: **AUC-ROC scores for encoder architectures on MIMIC-IV classification tasks.** Mean AUC-ROC scores with standard deviations (in parentheses) for each encoder evaluated on future disease, mortality, and rehospitalization prediction using a Random Forest classifier. Values in bold indicate the highest score per task.

KI-THRUST

On the KI-THRUST dataset, we compared our ICD2BERT model against the Pat2Vec models and the categorical encoder. We decided against running Med-BERT and BEHRT on this dataset due to the limited compute available and the results from MIMIC-IV and the following results on the EHRSHOT benchmark, which indicate little performance difference between the BERT models. Unexpectedly, the categorical encoder matches or exceeds BERT's performance on KI-THRUST, performing better at predicting future diseases and equal to BERT at predicting mortality and rehospitalization (Table 3). We decided to further investigate this result (see 'Future Disease Performance of Categorical Encoder' below). The KI-THRUST and, to a lesser extent, the MIMIC-IV dataset do not seem to benefit from the sequential ordering of ICD codes or from the attention mechanisms provided by BERT models. This may reflect the shorter average medical record lengths in these datasets compared to EHRSHOT, suggesting that transformer attention mechanisms provide less benefit when sequences are short. The performance of the categorical model is comparatively better on KI-THRUST, as this specific implementation was created for the

KI-THRUST[23] dataset. Also, given the performance on MIMIC-IV, it suggests that the information most relevant to these classification tasks is captured by the broader categorical structure of ICD codes, and that finer-grained code specificity contributes little additional signal. The Pat2Vec model performs the same as a random classifier on all tasks. It was not able to train effectively on the larger KI-THRUST dataset.

training dataset	model type	dimensionality	future_disease	mortality	rehospitalization
Ki-THRUST	BERT	240	0.70 (0.00)	**0.82 (0.00)**	**0.66 (0.00)**
MIMIC-IV	BERT	240	0.67 (0.00)	0.78 (0.00)	0.64 (0.00)
Ki-THRUST	BERT	768	0.63 (0.00)	0.70 (0.01)	0.63 (0.00)
MIMIC-IV	BERT	768	0.67 (0.00)	0.78 (0.00)	0.64 (0.00)
None	Categorical	241	**0.74 (0.00)**	**0.82 (0.00)**	0.65 (0.00)
Pat2Vec	Pat2Vec	10	0.50 (0.00)	0.50 (0.01)	0.50 (0.00)
Ki-THRUST	Pat2Vec	240	0.50 (0.00)	0.50 (0.01)	0.50 (0.00)
MIMIC-IV	Pat2Vec	240	0.63 (0.00)	0.70 (0.01)	0.61 (0.00)

Table 3: **AUC-ROC scores for encoder architectures on KI-THRUST classification tasks.** Mean AUC-ROC scores with standard deviations (in parentheses) for each encoder evaluated on future disease, mortality, and rehospitalization prediction using a Random Forest classifier. Values in bold indicate the highest score per task.

EHRSHOT / Few-Shot Learning

Across all four EHRSHOT task categories, BERT-based models consistently outperform all other architectures, with the categorical encoder placing second and Pat2Vec and CLMBR trailing behind (Fig. 2). Notably, comparing the CLMBR model's performance on our EHRSHOT benchmark, which is limited to ICD codes, to its performance on the full benchmark shows that removing other modalities increased the difficulty of all subtasks. This is especially apparent in the Chest-X-ray task, where the average model performance is barely above random, even on the full training data. This is an expected result, as it was also the hardest task in the original study, and the reduced input data exacerbated its difficulty. Furthermore, the missing data modalities seem to hinder the CLMBR model's performance more than those of other models. This is likely because CLMBR was pretrained on the full Stanford Healthcare feature set, in which ICD codes represent only a small fraction of the input tokens, making an ICD-only encoding a substantial distributional shift from its training conditions. The additional information used by the BEHRT and Med-BERT models does not improve their performance beyond that of the base BERT model. This is consistent with the results on the MIMIC-IV dataset, suggesting that the visit and age embeddings do not contribute useful information for the classification tasks. The Pat2Vec model performs near or at the bottom of all models with larger amounts of training data. However, it performs better than the CLMBR model on limited training data across all subtasks, and better than Med-BERT on Anticipating Lab Test Results. It falls below the CLMBR model across shot sizes 8-32, depending on the task, and below Med-BERT at shot size 24. This shows that Pat2Vec could be more data-efficient in

certain scenarios, delivering better results on small datasets. Its architecture, however, does not scale as well with growing dataset sizes. The categorical encoder's performance lies between that of the BERT models and the Pat2Vec and CLMBR models. It benefits less from the full training dataset than the BERT models do, as its performance increases the least when moving from 128 to all samples on the first two subtasks. This suggests that the intrinsic hierarchical structure of ICD codes captures meaningful information about a patient's medical history, though not as effectively as contextualized BERT encodings. Together, these results confirm that BERT architectures generalize effectively to low-data transfer settings and that the domain-specific modifications to BERT do not provide additional benefit in this benchmark.

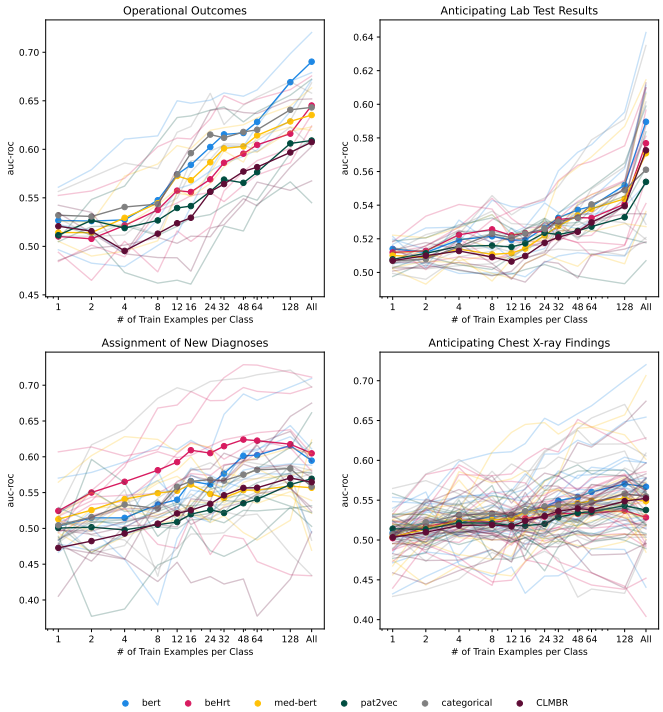


Fig. 2: **Few-shot learning performance on EHRSHOT dataset.** AUC-ROC scores for six encoder architectures across four task categories: Operational Outcomes, Anticipating Lab Test Results, Assignment of New Diagnoses, and Anticipating Chest X-ray Findings. Bold lines with filled markers show the mean performance across five runs for each encoder trained on MIMIC-IV; faint lines show individual task performance within each category.

Transfer Learning

From the above results, we also evaluate the impact of transferring BERT and Pat2Vec models between the MIMIC-IV and KI-THRUST datasets. For the BERT models, we observe a large drop in performance when transferring a model trained on KI-THRUST to MIMIC-IV, but only a small drop in the reverse direction. Because the KI-THRUST dataset contains only ICD-10 diagnosis codes, the model trained on this data cannot leverage the additional ICD categories in the MIMIC-IV dataset, resulting in lost information and reduced performance. In the reverse experiment, the number of categories decreases, indicating that the ICD-10 diagnosis codes convey similar information across both datasets. For the Pat2Vec models, the one trained on the MIMIC dataset outperforms the one trained on KI-THRUST in both cases. This suggests that the Doc2Vec architecture lacks the capacity to effectively model the larger KI-THRUST dataset.

Interpretability of ICD2BERT

A key advantage of ICD2BERT over domain-specific BERT models is its out-of-the-box compatibility with explainable AI packages for Transformers or BERT specifically. Because BERT is implemented using the standard Huggingface BERT architecture without additional input sequences, it is directly compatible with established tools without requiring any modifications. In contrast, Med-BERT and BEHRT require custom handling of their additional input sequences, which complicates the application of standard interpretability pipelines.

To demonstrate this compatibility, we applied SHAP values[31] to the ICD2BERT encoder-classifier pipeline on the MIMIC-IV mortality and rehospitalization tasks. Figure 3 shows force plots for correctly predicted cases, illustrating how individual ICD codes push the model prediction towards or away from a specific classification. For the rehospitalization example shown in Fig. 3 (bottom panel), the most influential codes supporting the prediction are d_10_E1165 and d_10_I10, and the codes against the prediction are p_10_0JB60ZZ and d_10_02100Z9. The two supporting codes have these descriptions: "Type 2 diabetes mellitus with hyperglycemia" and "Essential (primary) hypertension". The two opposing codes have these descriptions: "Excision of Chest Subcutaneous Tissue and Fascia, Open Approach" and "Mild hyperemesis gravidarum".[32] Interestingly, the

$$SEP$$

control token also has a large positive influence for this prediction, indicating that if all codes came from a single admission instead of two different ones, rehospitalization would be more likely. Being able to trace a prediction back to specific ICD codes in a patient's history is an important step toward building trust in model outputs, as it adds interpretability to an otherwise black-box model. Figure 4 shows the 10 most influential ICD codes in a beeswarm plot for mortality and rehospitalization prediction, respectively. Notably, the most influential codes are predominantly ICD-9 diagnosis codes, which is consistent with the ablation findings that ICD-9 codes are critical to MIMIC-IV model performance.

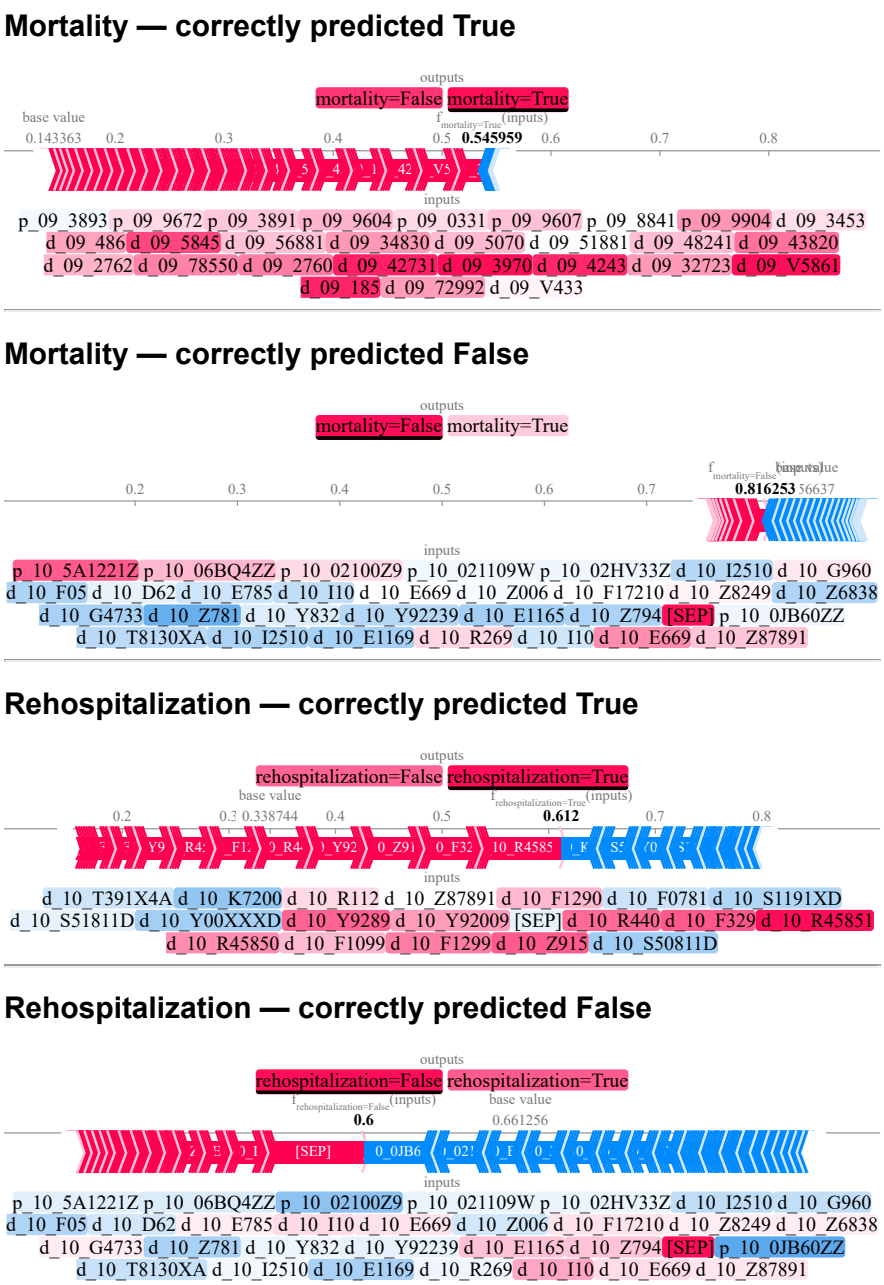
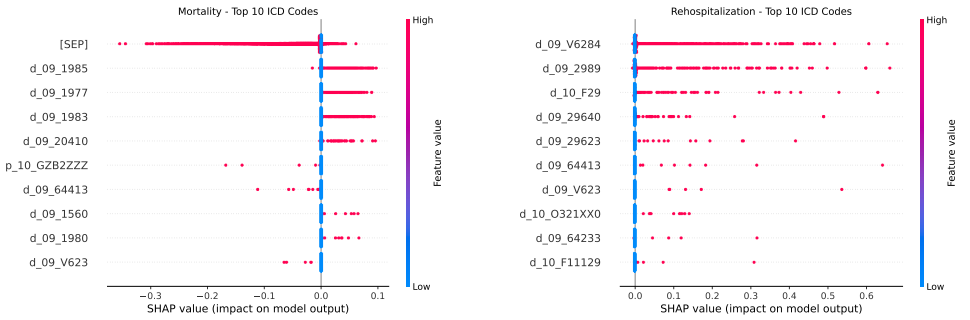


Fig. 3: **SHAP force plots on the MIMIC-IV.** Each plot shows how individual ICD codes push the model prediction towards a specific prediction. Red arrows and background indicate codes that increase the predicted probability, while blue indicates a decrease.



(a) (b)

Fig. 4: **Most influential ICD codes on MIMIC-IV** Left (a) shows SHAP values for mortality prediction and Right (b) shows values for rehospitalization prediction. The x-axis shows the SHAP value indicating the direction and magnitude of each code's impact on the predicted probability. Only codes with more than 5 occurrences are included.

Future Disease Performance of Categorical Encoder.

Due to the unexpectedly strong performance of the categorical encoder in predicting future diseases, we investigated this further. In the future disease prediction task, the model predicts whether a disease category was present at a patient's last admission. The categories correspond to the highest hierarchical level of an ICD-10 code, or, in other words, the first character of each code. The categorical encoding corresponds to the second-highest hierarchical level of the ICD codes, roughly the first two characters of each code. With such a task construction, there is a shortcut: map the hierarchy to the next level up. To evaluate whether the categorical model is taking this shortcut, we evaluated the Gini coefficients of a random forest classifier on the KI-THRUST dataset. We constructed a heatmap with the y-axis showing the input categories and the x-axis showing the output classifications. A diagonal line of high importance is visible in the corresponding feature importance heatmap (Supplementary Fig. 1), indicating that the random forest classifier used the classification shortcut explained above. Because past and current diseases correlate, this may explain the categorical model's strong predictive performance for future diseases.

In summary, the results so far show that the baseline categorical model is a very competitive choice on the MIMIC-IV and KI-THRUST datasets. However, the categorical encoder's strong performance on future disease prediction appears to be partially attributable to a hierarchical shortcut between the encoding structure and the task definition, rather than genuine clinical signal capture. Transferring a BERT model between datasets yields results only slightly worse than those of a natively trained model, as long as no new data modalities are present in the target dataset.

3 Discussion

In this study, we systematically compared six ICD code sequence encoding strategies across three datasets and found that domain-specific modifications to the BERT architecture do not meaningfully improve encoding performance relative to the unmodified base model. These results challenge a common assumption in the EHR encoding literature and suggest that the added complexity of models such as Med-BERT and BEHRT may not be justified for ICD-based classification tasks.

Our findings show that, when using the same training dataset, domain-specific additions to the BERT architecture do not improve performance. Specifically, the visit embeddings of the Med-BERT architecture and the age-embeddings of BEHRT provide little to no performance benefit while increasing model complexity and hindering interoperability. Including these additional input sequences complicates use with the existing ecosystem of deep learning and transformer packages, particularly XAI methods such as Shapley values, which expect a single 1D input sequence. To demonstrate this advantage, we applied SHAP values to ICD2BERT on the MIMIC-IV mortality and rehospitalization tasks. The resulting analysis identified ICD-9 diagnosis codes as the most influential features across both tasks, which is consistent with the ablation finding that ICD-9 codes are critical to MIMIC-IV model performance. The compatibility of ICD2BERT with standard SHAP implementations without modification demonstrates a practical benefit of avoiding domain-specific architectural changes. The model remains accessible to the full ecosystem of transformer interpretability tools, which is an important consideration for clinical deployment. Taken together, these results suggest that the primary driver of model performance is the quality and size of the pretraining dataset rather than domain-specific architectural adaptations.

The surprisingly competitive performance of the categorical encoder shows that it is not always necessary to build and train a complex, computationally intensive model for every task. Even though the categorical encoder works without contextualization and forgoes the sequential order of ICD codes, it still performs comparably to the BERT models. Especially noteworthy is its performance on the future disease classification task on the KI-THRUST dataset, where it surpassed the BERT model in AUC-ROC score. After further analysis, we found that this particular task favored the categorical encoder. We analyzed the trained classifier and found that it was mapping each of the encoder's categories onto a consistent output category. These findings indicate that the categorical encoder's strong performance on future disease prediction should be interpreted with caution. It appears to exploit a structural shortcut rather than capture genuine longitudinal clinical patterns. Still taking into account the model's performance on predicting rehospitalization and mortality on MIMIC-IV and KI-THRUST, and on all EHRSHOT tasks, it can compete with the BERT models. While it is still outperformed slightly by the BERT models in these cases, it trades this performance for simplicity, no need for large training data, and interpretability.

Analyzing the information contained in the ICD codes revealed three characteristics of the MIMIC-IV dataset. Firstly, the majority of the information is contained in the diagnosis codes, as the BERT model trained only on them did not suffer a performance hit. Secondly, the ICD-9 diagnosis codes are still important for the classification tasks as discarding them resulted in the lowest performing models in all tasks.

Thirdly, the latter characters of the ICD codes, which increase the specificity of the diagnosis, do not contain useful information for our classification tasks. The models that used only the first 2 or 3 characters of the ICD codes did not show a decrease in performance due to the reduction in specificity.

We next asked whether these MIMIC-IV findings held under transfer. When transferring the BERT models trained on the ablated MIMIC-IV dataset to KI-THRUST, we found that including ICD-9 codes during training did not affect evaluation performance. As KI-THRUST only contains ICD-10 codes, training the model on ICD-9 codes did not improve performance, but it also did not hinder it. The results when reducing the specificity of ICD codes differ from the MIMIC-IV dataset. On KI-THRUST, the performance of BERT models restricted to 2- and 3-character code lengths decreases drastically compared to using the full ICD codes. We find two possible explanations for this. Either the latter characters of ICD codes contain a significant amount of information in the KI-THRUST dataset, or while individual full ICD codes carry similar meaning across both datasets, the coarser categories formed by truncating to 2–3 characters group different underlying conditions together in KI-THRUST than in MIMIC-IV.

Transfer performance mirrored the code-coverage logic of our ablation results: models pretrained on MIMIC-IV's broader mix of code types transferred well to KI-THRUST, but models trained only on KI-THRUST's ICD-10 diagnoses could not perform well when applied to MIMIC-IV. Evaluating a BERT model on KI-THRUST while training it on MIMIC-IV showed a small but measurable decrease in performance. Notably, the model has the same AUC-ROC scores on all three tasks as the model trained only on ICD-10 diagnosis codes on MIMIC-IV. Doing the transfer in reverse resulted in a steeper drop-off, making the model perform the same as one trained only on ICD-10 codes in MIMIC-IV. This shows that for transfer learning to be successful, the training dataset should include all modalities present in the target dataset. Additional modalities beyond what the target requires did not improve transfer performance, but did not hurt it either, so training on the broadest available modality set maximizes the number of datasets a model can later transfer to, at no apparent cost.

Analyzing the results of the Pat2Vec models revealed that the architecture is largely superseded by BERT for ICD code encoding. Starting with the pretrained model from the original paper, it was not able to encode meaningful information about the patient's medical history. We suspect this is largely due to the 10-dimensional encoding being too small to capture all the information in a medical history. The transfer of the model to a new dataset also adds to this challenge, but it does not explain why the model performs poorly on all tasks. The models trained on MIMIC-IV and KI-THRUST deliver decent results; however, they are below those of the BERT models. Unexpectedly, the model trained on KI-THRUST performed poorly on its original dataset, yet surpassed the model trained on MIMIC-IV in predicting future diseases on MIMIC-IV. This remains unexplained and warrants further investigation. Transferring the Pat2Vec model from MIMIC-IV to EHRSHOT yielded strong results in low-data scenarios, where Pat2Vec surpasses or matches some BERT models at low shot sizes. The model loses out to those models with increasing shot sizes. Together

with the results on MIMIC-IV and KI-THRUST, this indicates that Pat2Vec has stability or performance issues on large datasets. It performs worse than BERT models as the dataset size increases and behaves unexpectedly when trained on our largest KI-THRUST dataset.

A major limitation of our study is the smaller-scale analysis on the KI-THRUST dataset. Due to computation constraints, we limited the number of models evaluated on this dataset, leaving out Med-BERT, BEHRT, and the training of ablation models. The results of MIMIC-IV and EHRSHOT indicate that the performance difference would be negligible. This study is limited to ICD codes as the data modality, potentially underutilizing the transformer architecture's capabilities. As seen with the pretrained CLMBR model, its performance is severely reduced when evaluated only on ICD codes rather than across all its data modalities. While MIMIC-IV and KI-THRUST contain enough data to train a solid pretrained BERT model, they are more than an order of magnitude smaller than the dataset used to pretrain Med-BERT. The KI-THRUST dataset spans only 1 year of records, preventing or limiting the capture of long-term disease trajectories. The categorical encoder's shortcut finding may be specific to the hierarchical structure of our future disease prediction task and may not generalize to other task definitions.

Future work should address these limitations by incorporating additional data modalities alongside ICD codes. This may particularly benefit BERT models, given their capacity to handle longer input sequences. Training and evaluating on larger datasets with potentially longer temporal coverage would clarify whether the sequential and contextual encodings of BERT become more pronounced over extended medical histories. Additionally, the unexpected data efficiency of Pat2Vec encodings in low-shot settings and its unexpected performance on larger datasets warrant further systematic investigation, as it may offer practical advantages in resource-constrained deployment scenarios.

In summary, domain-specific modifications to BERT do not improve performance in ICD code sequence encoding. The primary drivers of model performance are the quality and size of the pretraining dataset, not architectural complexity. Furthermore, the categorical encoder remains competitive across most tasks. Practitioners should therefore carefully weigh the computational cost of transformer-based models against the simplicity and interpretability of baseline approaches before committing to more complex architectures.

4 Methods

4.1 Data

We evaluated transformer-based encoders across three different biomedical datasets that differ in coding systems, population size, and data granularity. The MIMIC-IV database[21] was used to assess model behavior across different International Classification of Diseases (ICD) versions. It comprises de-identified electronic health records from critical care units at the Beth Israel Deaconess Medical Center, containing both ICD-9 and ICD-10 diagnostic codes, which allowed systematic evaluation of model

robustness to code version heterogeneity. The second dataset, KI-THRUST[23], contains large-scale German routine insurance claims data from hospitals and outpatient providers. It includes only ICD-10 codes but covers a substantially broader and more heterogeneous population, providing a valuable benchmark for large-cohort generalization. Finally, the EHRSHOT[22] dataset consists of a smaller cohort of patients with extensive longitudinal records and dense visit histories. This dataset was used to evaluate the transfer learning and fine-tuning capabilities of pretrained transformer encoders in settings where long-term temporal context and detailed individual trajectories are available.

For MIMIC and KI-THRUST, we compiled all ICD codes for each patient, grouping them by admission and, when possible, sorting them by time of entry. All of a patients' admissions were then grouped, leaving the last one out for the prediction tasks. The codes from the last admission were binarized by the first letter of the ICD code, resulting in 25 categories, every letter of the alphabet excluding U. Each category was treated as a binary classification problem, with patients' previous ICD codes as inputs and the presence of a code beginning with the corresponding letter in their final admission as the target. For rehospitalization prediction, the target was a binary classification indicating whether the patient's final admission occurred within 30 days of the penultimate admission. For the mortality prediction task, the target was a binary flag in the final admission indicating whether the patient died during or soon after the admission.

For the EHRSHOT Dataset, only entries containing ICD codes were kept for each patient. To select the ICD codes relevant to an individual classification, all ICD codes for the patient prior to the classification date were collected and combined using the same method as in the MIMIC and KI-THRUST Dataset. We used the same few-shot evaluation method as the original paper. The original few-shot examples could not be used because some patients had no ICD code entries prior to the classification timestamp. The few-shot examples were reselected from classification tasks in which patients had ICD codes in their medical history.

4.2 Experiments

Experimental Baseline

Both the MIMIC and KI-THRUST datasets were split into training, validation, and test sets. MIMIC was split with ratios 70-15-15 and KI-THRUST was split 64-16-20 which is the results of a predefined 20% test split on the KI-THRUST dataset and the remaining data being split 4/5 to 1/5 for the training and validation set. For EHRSHOT, their predefined train-validation-test split was used. The encoders were then either trained on the MIMIC or the KI-THRUST datasets. All ICD sequences were then encoded and used as input to either a Logistic Regression or a Random Forest classifier. For statistical purposes, the training data for the classifiers were resampled 10 times via bootstrapping for the MIMIC and KI-THRUST datasets, yielding 10 estimates of the encoding performance. For the EHRSHOT dataset, the training and validation sets were used to randomly select samples for the few-shot test 5 times, mimicking the few-shot evaluation setup of the original paper. The classifiers' hyperparameters (see Supplementary for details) were optimized using a grid

search, and the best-performing model on the validation set was evaluated on the test set. The model selection metric used was the AUC-ROC score, as it performs well on imbalanced datasets, which are common in medical datasets and are also present in the datasets used for this study.

For the experiments, one encoder of each type was trained on each of the two training datasets and subsequently evaluated across all three datasets to assess both in-distribution performance and transfer learning capabilities.

BERT Architecture Experiments

To evaluate the possible impact of additional metadata on the encoding performance of BERT models, we compared our ICD2BERT model with two popular Models. These models extend the BERT architecture by incorporating additional data for each ICD code, such as the patient's age at the time the code was entered. Specifically, we chose the Med-BERT[26] architecture, which includes the patient's current visit count at the time of ICD code entry, and the BEHRT[25] architecture, which includes the patient's age at the time of ICD code entry. We compared both implementations to a standard BERT model that uses only the ICD codes as input.

BERT Ablation Experiments

For the final experiments, we analyzed the usefulness of each ICD code category and the specificity of ICD codes. We modified the MIMIC dataset to create three reduced training datasets. The first dataset contains only ICD-10 codes to determine whether data on ICD-9 codes can assist with model training. The second dataset contains only diagnosis codes to see whether procedure codes provide additional information about a patient or if all information is already contained in the diagnosis codes. The final dataset contains only ICD-10 diagnosis codes to evaluate the effect of omitting both the older ICD revisions and potentially superfluous procedure codes.

Finally, we wanted to test the effect of reducing ICD code specificity, as part of the approaches used in other papers for encoding ICD sequences. To test this hypothesis, we constructed three new datasets from the MIMIC database. Since ICD Procedure codes do not conform to a hierarchical structure, we conducted these tests only on ICD diagnosis codes. We employed two different methods to generalize the ICD codes. For the first approach, we used the categories defined by the categorical encoder and mapped ICD codes to their corresponding categories, resulting in a vocabulary of 241 ICD codes. For the second approach, we leveraged the structure of ICD diagnosis codes, in which each letter increases specificity. We limited the maximum length of ICD codes to 2 or 3 characters to create two more test cases with 351 and 2952 different ICD codes as the vocabulary. With these two testing strategies, we evaluated the viability of generalizing ICD codes using categorizations created by clinical experts, as well as a simple generalization that leverages the intrinsic structure of ICD codes.

4.3 Model Training

All Encoders were trained either on the MIMIC or KI-THRUST training data split. The model's output dimensionality was set to 240 to ensure a fair comparison with the baseline clinical categories, which have 241 output values. The BERT models used

masked language modeling for training. The Pat2Vec models were trained using either of the two methods described in the Doc2Vec paper: distributed memory or distributed bag-of-words. The Pat2Vec hyperparameters, including the training method, were optimized using the same parameter ranges and optimization algorithm as the original paper. The categorical encoder does not require training as it utilizes only the ICD code categories created by clinicians. We evaluated two variants: a 241-dimensional encoder covering only ICD-10 diagnosis categories, matching the scope of the other encoders' primary comparison, and a 640-dimensional encoder additionally covering ICD-9 and ICD-10 procedure codes as well as ICD-9 diagnosis codes, matching the full code coverage of the MIMIC-IV dataset (Table 1). Each entry indicates whether a given ICD code category is present in the patient's history. The CLMBR model used pretrained weights without further training or finetuning.

The classifiers are trained on encodings for each classification task, using a grid search to optimize hyperparameters and maximize the AUC-ROC. The best model from the grid search is then evaluated on the test set.

Each classifier is trained 10 times on the MIMIC and KI-THRUST datasets using bootstrapped training sets. The few-shot training on the EHRSHOT dataset is repeated 5 times across different patients, and the full-dataset evaluation is repeated 5 times with different classifier initializations.

4.4 ICD Encodings

To create input data for model training and subsequent classification tasks, the ICD codes were translated into a format that prevents codes from different revisions or diagnosis and procedure codes from being mistaken for one another. The MIMIC dataset has both ICD-9 and ICD-10 codes as well as diagnosis and procedure codes. Due to the preprocessing of the ICD codes in this dataset, some codes had the same name but different meanings across multiple categories. Each code was prefixed with either "d" for diagnosis codes or "p" for procedure codes, followed by the ICD revision number. These prefixes were then attached to the ICD codes and separated by underscores, resulting in the ICD-10 diagnosis code **I10.9** being translated to **d_10_I109**.

4.5 Explainable AI

The value of a model in medicine is not only determined by its performance but also by its transparency and interpretability. For this reason, we chose the Random Forest Classifier. This model is intrinsically explainable, which enables us to analyze the model itself for interpretability as well as use end-to-end interpretability methods to evaluate the encoder classifier pipeline as a whole.

To evaluate the classifiers, we analyzed the feature importances generated by the Random Forest Classifier during training. For the categorical encoder, where each feature corresponds to a clinically defined ICD category, this provides directly interpretable insight into which code categories most influence each classification task. A shortcoming of this strategy is that the input data must have an interpretable meaning itself, which is not the case for BERT or Pat2Vec models. For global and patient-level interpretability of our ICD2BERT model, we used SHAP values[31] combined

with a Random Forest Classifier trained on MIMIC-IV's rehospitalization and mortality prediction tasks. Because ICD2BERT uses the standard Hugging Face BERT architecture without additional input sequences, it is directly compatible with existing SHAP implementations without requiring any modifications. SHAP values were computed for each ICD code in each patient's history, yielding both patient-level force plots that explain individual predictions and global beeswarm plots that summarise the most influential codes across the dataset. The most influential codes were determined by their mean SHAP value, and only codes with more than 5 occurrences were considered. For the singular examples, samples were randomly selected from correctly classified patients for each task and outcome. Only patients with 15-30 ICD codes were included in the singular examples, purely for presentation purposes.

4.6 Implementation Details

Full implementation details, including hyperparameters, software versions, and hardware specifications, are provided in Supplementary Methods. Briefly, ICD2BERT was initialized from the pretrained bert_base_uncased weights with a replaced vocabulary and reduced hidden dimensionality (240). All models were trained using standard optimization settings consistent with their original publications.

4.7 Ethics

All methods were carried out in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki, the EU General Data Protection Regulation, and the German Social Code provisions governing social data (§ 35 SGB I, §§ 67 and 75 SGB X, §§ 284–287 SGB V). Analysis of the KI-THRUST data was conducted within the project KI-THRUST, funded by the Innovation Committee of the Federal Joint Committee (G-BA; grant 01VSF20014) under § 92a (2) SGB V. The study was approved by the Ethics Committee of the Lower Saxony Medical Association (Ethikkommission der Ärztekammer Niedersachsen; positive vote of 7 February 2022). Transfer of the statutory health and long-term care insurance data for research purposes was approved by the German Federal Office for Social Security (Bundesamt für Soziale Sicherung, BAS) on 3 June 2022 under § 75 SGB X, including review and approval of the project's data protection concept. The requirement for informed consent was waived under § 75 SGB X, which permits the transfer of social data for research purposes without the consent of the data subjects where the data are indispensable to the research and the public interest in the research substantially outweighs the private interest of the individuals concerned; this waiver was granted by the BAS as part of the approval. All data were pseudonymized prior to transfer to the researchers. MIMIC-IV and EHRSHOT are de-identified datasets used under their respective data use agreements.

Declarations

Acknowledgements. We gratefully acknowledge the support of the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003).

The authors gratefully acknowledge the computing time provided to them on the high-performance computer Lichtenberg II at TU Darmstadt, funded by the German Federal Ministry of Education and Research (BMBF) and the State of Hesse.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 567400630

The authors gratefully acknowledge the computing time granted by the KISSKI project. The calculations for this research were conducted with computing resources under the project kisski-umg-fairpact-2.

The Innovation Committee of the Federal Joint Committee granted this study under the funding reference 01VSF20014 (KI-THRUST)

Data availability. The MIMIC-IV dataset is available to credentialed researchers via PhysioNet at https://physionet.org/content/mimiciv/, subject to completion of the required training and a data use agreement. The EHRSHOT dataset is available at https://stanford.redivis.com/datasets/6aj5-05w5awaxd, subject to a Stanford research data use agreement. The KI-THRUST data are not publicly available. They are pseudonymized claims data of German statutory health insurance funds (BAHN-BKK, Novitas BKK, Pronova BKK, Siemens-Betriebskrankenkasse), transferred to the aQua-Institut under a project-specific authorization and may not be shared or redistributed by the authors. Requests regarding these data should be directed to the aQua-Institut (office@aqua-institut.de).

Code availability. To enable reproducibility and to benefit other EHR-based studies, we share our source code on GitHub https://github.com/HauschildLab/ICD2BERT. The pretrained ICD2BERT model is available on Huggingface https://huggingface.co/PredLMed/ICD2BERT under the same license as the MIMIC-IV dataset.

References

[1] World Health Organization: International Statistical Classification of Diseases and Related Health Problems: 10th Revision (ICD-10). World Health Organization

[2] Parasrampuria, S., Henry, J.: Hospitals' use of electronic health records data, 2015–2017 **46**(1), 13

[3] World Health Organization: Importance of ICD. World Health Organization. https://www.who.int/standards/classifications/classification-of-diseases/who-family-of-international-classifications/importance-of-icd

[4] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y.: Artificial intelligence in healthcare: past, present and future **2**(4) https://doi.org/10.1136/svn-2017-000101

[5] Kruse, C.S., Stein, A., Thomas, H., Kaur, H.: The use of electronic health records to support population health: A systematic review of the literature **42**(11), 214 https://doi.org/10.1007/s10916-018-1075-6

[6] Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A.B., Kronzer, A.: Predicting hospital readmission via cost-sensitive deep learning **15**(6), 1968–1978 https://doi.org/10.1109/TCBB.2018.2827029

[7] Naser, P.V., Maurer, M.C., Fischer, M., Karimian-Jazi, K., Ben-Salah, C., Bajwa, A.A., Jakobs, M., Jungk, C., Jesser, J., Bendszus, M., Maier-Hein, K., Krieg, S.M., Neher, P., Neumann, J.-O.: Deep learning aided preoperative diagnosis of primary central nervous system lymphoma. iScience **27**(2), 109023 (2024) https://doi.org/10.1016/j.isci.2024.109023

[8] Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare **2**(10), 719–731 https://doi.org/10.1038/s41551-018-0305-z

[9] Shen, J., Zhang, C.J.P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P.-H., Ming, W.-K.: Artificial intelligence versus clinicians in disease diagnosis: Systematic review **7**(3), 10010 https://doi.org/10.2196/10010 . Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics

[10] Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C.S., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., He, L., Zhu, J., Tian, P., Shao, H., Zheng, L., Hou, R., Hewett, S., Li, G., Liang, P., Zang, X., Zhang, Z., Pan, L., Cai, H., Ling, R., Li, S., Cui, Y., Tang, S., Ye, H., Huang, X., He, W., Liang, W., Zhang, Q., Jiang, J., Yu, W., Gao, J., Ou, W., Deng, Y., Hou, Q., Wang, B., Yao, C., Liang, Y., Zhang, S., Duan, Y., Zhang, R., Gibson, S., Zhang, C.L., Li, O., Zhang, E.D., Karin, G., Nguyen, N., Wu, X., Wen, C., Xu, J., Xu, W., Wang, B., Wang, W., Li, J., Pizzato, B., Bao, C., Xiang, D., He, W., He, S., Zhou, Y., Haw, W., Goldbaum, M., Tremoulet, A., Hsu, C.-N., Carter, H., Zhu, L., Zhang, K., Xia, H.: Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence **25**(3), 433–438 https://doi.org/10.1038/s41591-018-0335-9

[11] Miotto, R., Weng, C.: Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials **22**, 141–150

[12] Richens, J.G., Lee, C.M., Johri, S.: Improving the accuracy of medical diagnosis with causal machine learning **11**(1), 3923 https://doi.org/10.1038/s41467-020-17419-7

[13] Shahid, A.H., Khattak, W.A.: Improving patient care with machine learning: A

game-changer for healthcare **5**(1), 150–63

[14] Casey, J.A., Schwartz, B.S., Stewart, W.F., Adler, N.E.: Using electronic health records for population health research: A review of methods and applications **37**, 61–81 https://doi.org/10.1146/annurev-publhealth-032315-021353

[15] Tran, T., Luo, W., Phung, D., Gupta, S., Rana, S., Kennedy, R.L., Larkins, A., Venkatesh, S.: A framework for feature extraction from hospital medical data with applications in risk prediction **15**(1), 425 https://doi.org/10.1186/s12859-014-0425-8

[16] Azarm-Daigle, M., Kuziemsky, C., Peyton, L.: A review of cross organizational healthcare data sharing **63**, 425–432 https://doi.org/10.1016/j.procs.2015.08.363

[17] Weiskopf, N.G., Weng, C.: Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research **20**(1), 144–151 https://doi.org/10.1136/amiajnl-2011-000681

[18] Weiskopf, N.G., Hripcsak, G., Swaminathan, S., Weng, C.: Defining and measuring completeness of electronic health records for secondary use **46**(5), 830–836 https://doi.org/10.1016/j.jbi.2013.06.010

[19] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives **35**(8), 1798–1828

[20] Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care **13**(6), 395–405 https://doi.org/10.1038/nrg3208

[21] Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.-w.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset **10**(1), 1 https://doi.org/10.1038/s41597-022-01899-x

[22] Wornow, M., Thapa, R., Steinberg, E., Fries, J.A., Shah, N.H.: EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2307.02028

[23] Gemeinsamer Bundesausschuss (G-BA): KI-THRUST – Potenziale KI-gestützter Vorhersageverfahren auf Basis von Routinedaten. G-BA Innovationsfonds. https://innovationsfonds.g-ba.de/projekte/neue-versorgungsformen/ki-thrust.559

[24] Steiger, E., Kroll, L.E.: Patient embeddings from diagnosis codes for health care prediction tasks: Pat2vec machine learning framework **2**(1), 40755 https://doi.org/10.2196/40755 . Company: JMIR AI Distributor: JMIR AI Institution: JMIR AI Label: JMIR AI

[25] Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: BEHRT: Transformer for electronic health records **10**(1), 7155 https://doi.org/10.1038/s41598-020-62922-y

[26] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction **4**(1), 86 https://doi.org/10.1038/s41746-021-00455-y

[27] Shang, J., Ma, T., Xiao, C., Sun, J.: Pre-training of Graph Augmented Transformers for Medication Recommendation. arXiv. https://doi.org/10.48550/arXiv.1906.00346

[28] NHS England Digital: Read Codes. NHS England Digital. https://digital.nhs.uk/services/terminology-and-classifications/read-codes

[29] Kuan, V., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Bhatti, O., Husain, S., Sutaria, S., Hingorani, M., Nitsch, D., Parisinos, C.A., Lumbers, R.T., Mathur, R., Sofat, R., Casas, J.P., Wong, I.C.K., Hemingway, H., Hingorani, A.D.: A chronological map of 308 physical and mental health conditions from 4 million individuals in the english national health service **1**(2), 63–77 https://doi.org/10.1016/S2589-7500(19)30012-3

[30] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196. PMLR

[31] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017)

[32] Centers for Disease Control and Prevention: ICD-10: International Classification of Diseases, Tenth Revision. https://www.cdc.gov/nchs/icd/icd-10.htm

Author Contributions

JH is the first Author. AH initialized the conceptualization of the project. JH led the design, implementation, and execution of the study with input from AH. JM, MM, and ZR developed the baseline model and evaluation schema on the KI-THRUST data. LW, TG, TP, and MK preprocessed and managed the KI-THRUST data. All authors contributed to writing and proofreading the manuscript.

Competing Interest

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf