Attention Networks for Augmenting Clinical Text with Support Sets for Diagnosis Prediction

Paul Grundmann*, **Tom Oberhauser***, **Felix Gers and Alexander Löser**
Berliner Hochschule für Technik
Luxemburger Str. 10, 13353 Berlin
{pgrundmann, toberhauser, gers, aloeser}@bht-berlin.de
**Both authors contributed equally.*

Abstract

Diagnosis prediction on admission notes is a core clinical task. However, these notes may incompletely describe the patient. Also, clinical language models may suffer from idiosyncratic language or imbalanced vocabulary for describing diseases or symptoms. We tackle the task of diagnosis prediction, which consists of predicting future patient diagnoses from clinical texts at the time of admission. We improve the performance on this task by introducing an additional signal from support sets of diagnostic codes from prior admissions or as they emerge during differential diagnosis. To enhance the robustness of diagnosis prediction methods, we propose to augment clinical text with potentially complementary set data from diagnosis codes from previous patient visits or from codes that emerge from the current admission as they become available through diagnostics. We discuss novel attention network architectures and augmentation strategies to solve this problem. Our experiments reveal that support sets improve the performance drastically to predict less common diagnosis codes. Our approach clearly outperforms the previous state-of-the-art PubMedBERT baseline by up 3% points. Furthermore, we find that support sets drastically improve the performance for pregnancy- and gynecology-related diagnoses up to 32.9 % points compared to the baseline.

1 Introduction

Pre-trained large language models such as ClinicalBERT (Alsentzer et al., 2019) or PubMedBERT (Gu et al., 2021) are commonly used in the medical domain to predict diagnoses from admission notes (Hashir and Sawhney, 2020; Sushil et al., 2018a; van Aken et al., 2021). Admission and discharge notes are valuable information sources about doctors' decisions about patients and the outcomes. However, the vocabulary in these notes is often insufficient to describe the patients' clinical phenotype fully. Also, clinical text frequently contains idiosyncratic vocabulary, uncommon abbreviations and differs from clinic to clinic in writing style. Moreover, evidence for clinical diseases in the text is imbalanced. In particular for less common or even rare diseases (see also Figure 1) not much text evidence exists. Finally, pre-trained language models can suffer from limited access to training data because of silos or data-privacy concerns. These factors can lead to poor performance in predicting outcomes on clinical text with pre-trained language models.

Multimodal Patient Representation. Miotto et al. (2016) and Topol (2019) therefore propose to augment text with potential complementary multimodal data into a reusable *deep patient representation* to improve clinical prediction tasks. For example, recent work surveys to augment text with image data (Esteva et al., 2021), with complementary medical text books (van Aken et al., 2021), ontologies (Cai et al., 2020) or time series data (Yang and Wu, 2021).

Improving predictions with text and set data. A particularly powerful source to augment clinical text are sets of diagnosis codes from previous visits of the patient or upcoming hypotheses of the treating physician during the patients' treatment. For example, a patient in an ICU scenario receives on average a set of more than ten diagnosis codes at discharge time, see also Table 2. These sets match a patients' previous state against a common ontology, such as ICD or CCS medical nomenclature. Therefore, these sets are a rich and potentially complementary knowledge source for a patient representation. To our best knowledge this is the first work on augmenting clinical text for diagnosis prediction with such sets. Figure 2 illustrates this novel task in detail: Given is the admission note containing details on chief complaints, present illness, medication, physical examination and family or social history. In addition, the system receives a

Proceedings of the 29th International Conference on Computational Linguistics, pages 4765–4775
October 12–17, 2022.

support set of additional diagnosis codes observed for this patient in the past or during the current treatment. Given both inputs, the final task is to predict the likely diagnostic outcome for the patient at discharge time.

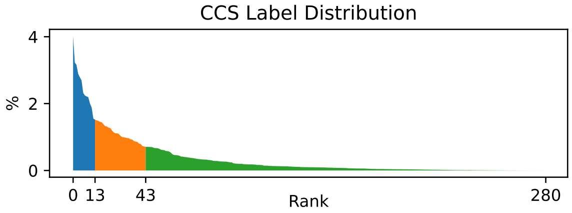


Figure 1: The distribution of CCS labels of the full dataset. The top 13 codes form the short head of the distribution. The middle tail is 30 codes wide, while the remaining 237 codes form the long tail.

Contribution. Augmenting text with set data is a complex knowledge integration problem. Ideally, a multi-modal representation from both, text and set data, is much more powerful for large multi-label classification tasks, such as diagnosis prediction, than each knowledge representation on its own. For solving this problem, our major contributions are: **(1)** We represent text and set data in two different latent vector spaces. This also includes investigating different sampling methods of set elements for learning embeddings during training. Optionally, we enrich disease codes in sets with additional textual information from UMLS, a medical ontology, and Wikidata. **(2)** We propose three different novel network architectures for augmenting knowledge from text with set data, including pooled- and full text attention as well as a dual stack encoder. **(3)** In a rigid experimental setting, we compare these architectures against each other and two strong baselines. We also report prediction results in particular for infrequent diseases on the MIMIC-III data set with approximately 60.000 admissions and more than 2 million clinical notes. The remainder of this paper is structured as follows: We review related work in Section 2. In Section 3, we explore task and data set characteristics, from which we justify in Section 4 our novel network architecture design. Section 5 reports our quantitative evaluation, followed by result discussion and an error analysis in Section 6. Finally, we conclude in Section 7.

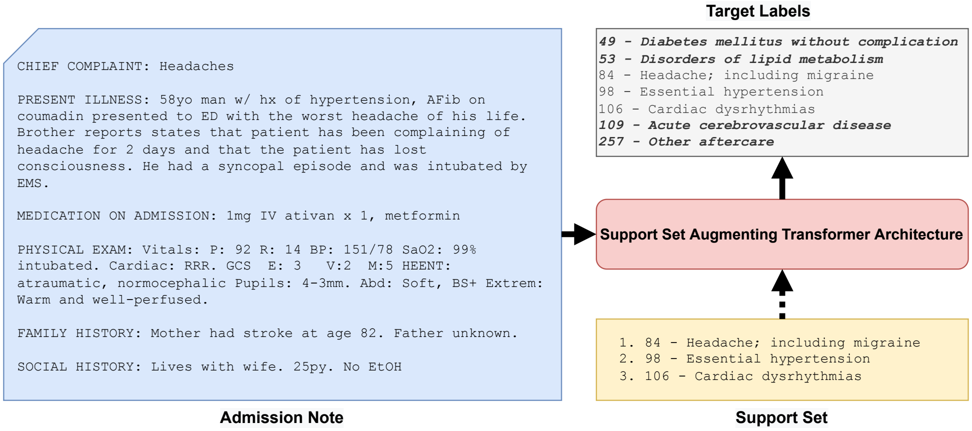
2 Related Work

There is a large amount of work focusing on diagnosis prediction from EHR data, especially clinical codes. Furthermore, there is an increasing emphasis on incorporating text or multi-modal data from clinical notes. We distinguish ourselves, particularly from work in ICD coding, since only information at the time of admission is used for our considered tasks. ICD coding, on the other hand, uses all data available at discharge time.

Diagnosis prediction on codes. Choi et al. (2016) use a reverse time attention mechanism on the diagnosis and procedure codes of the patients' history for the task of heart failure prediction. Ma et al. (2017) use a bidirectional recurrent neural network (RNN) on the diagnosis and procedure codes of the patients' history to predict diagnosis codes for the next admission. Later they apply graph-based attention to incorporate the knowledge of a medical knowledge graph to learn medical representations (Ma et al., 2018). Peng et al. (2020) use a self-attention mechanism on the diagnosis codes to capture contextual and temporal relations within the patients' journey to predict the second hierarchy of the ICD-9 codes.

Clinical text for diagnosis prediction. Boag et al. (2018) evaluate the usefulness of different simple text representations for diagnosis prediction and show that the text itself contains valuable information. Sushil et al. (2018b) use stacked denoising autoencoders combined with a paragraph vector model to learn patient representations. van Aken et al. (2021) simulate patients at admission time by only using parts of the textual descriptions known at admission time, such as *"Chief complaint"* or *"Medical history"*. Winter et al. (2022) apply knowledge graphs to retrain and instill attention heads with complementary structured domain knowledge for clinical outcome prediction from text. Papaioannou et al. (2022) seek to embed complementary knowledge to increase the performance on low resource languages by consecutive fine-tuning of multi-lingual models.

Multimodal diagnosis prediction. Lipton et al. (2016) use an LSTM architecture on 13 time-series variables like blood pressure or heart rate. Liu et al. (2018) use free, unstructured text from medical notes and structured clinical information such as numerical lab and vital sign values to predict a small set of specific chronic diseases. Qiao et al. (2019) use RNNs and attention to mix code- and text features for readmission diagnosis prediction from prior admissions. In contrast to the aforemen-


Target Labels
49 - Diabetes mellitus without complication
53 - Disorders of lipid metabolism
84 - Headache; including migraine
98 - Essential hypertension
106 - Cardiac dysrhythmias
109 - Acute cerebrovascular disease
257 - Other aftercare
Support Set Augmenting Transformer Architecture
1. 84 - Headache; including migraine
2. 98 - Essential hypertension
3. 106 - Cardiac dysrhythmias
Admission Note
Support Set

```
CHIEF COMPLAINT: Headaches

PRESENT ILLNESS: 58yo man w/ hx of hypertension, AFib on
coumadin presented to ED with the worst headache of his life.
Brother reports states that patient has been complaining of
headache for 2 days and that the patient has lost
consciousness. He had a syncopal episode and was intubated by
EMS.

MEDICATION ON ADMISSION: 1mg IV ativan x 1, metformin

PHYSICAL EXAM: Vitals: P: 92 R: 14 BP: 151/78 SaO2: 99%
intubated. Cardiac: RRR. GCS  E: 3   V:2  M:5 HEENT:
atraumatic, normocephalic Pupils: 4-3mm. Abd: Soft, BS+ Extrem:
Warm and well-perfused.

FAMILY HISTORY: Mother had stroke at age 82. Father unknown.

SOCIAL HISTORY: Lives with wife. 25py. No EtOH
```

Figure 2: Illustration of the set-augmented diagnosis prediction task: Given an admission note and an optional support set of diagnosis codes, our architecture aims to predict the diagnostic outcome for the patient at discharge time. Our basic observation is that the text data is only partially representing the patient, might be noisy and diagnostic codes can complement observations from the clinical text. Combining information from both sides will lead to improved understanding of a latent model towards clinical health conditions.

tioned approaches, we aim to model interdependencies between diagnostic codes and mix them with state-of-the-art text representations of clinical admission notes. Further, we seek to implement an interactive system that allows verifying hypotheses by systematically adding or removing diagnostic codes in combination with a current admission note.

Distinction from previous work. In contrast to the aforementioned approaches, we aim to model interdependencies between diagnostic codes and mix them with state-of-the-art text representations of clinical admission notes. Further, we seek to implement an interactive system that allows verifying hypotheses by systematically adding or removing diagnostic codes in combination with a current admission note. Our model does not rely on the existence of codes but instead uses them as they become available during a diagnostic process or through prior admissions to refine the classification result. Up to our knowledge, there is only related work that uses RNNs / LSTMs (Qiao et al., 2019) for representing the set embeddings. Modeling the set embeddings with RNNs is problematic because they introduce temporal dependencies between the diagnosis code inputs. In consequence, they treat the diagnosis code sets as a sequence. Most of the time, those temporal dependencies are not reflected in the available data.

3 Tasks and Datasets

In the following section, we introduce the tasks of diagnosis and readmission diagnosis prediction and describe our medical dataset.

Prediction from clinical text and diagnosis sets. Following van Aken et al. (2021), our model aims to predict diagnostic codes assigned to a patients' admission after their discharge with the constraint of using only information available at admission time [1]. In addition, we allow the model to leverage an optional set of support codes to refine the classification process. In the real-world use of our model, these supporting codes would originate from a doctors' hypotheses, evident diagnoses or from a diagnosis of a doctor outside the clinic the patient has visited before, such as the family doctor. Formally, our training data consists of a set of admissions $\mathbb{A}$ where $A_i = (T_i, S_i, C_i)$, $A_i \in \mathbb{A}$. $T_i = (t_1, \ldots, t_n)$ is the text of the admission note with a sequence length of n tokens for a patient at admission time. $C_i \subset \mathbb{C}$ is the prediction target of diagnostic codes from the label space $\mathbb{C}$ and $S_i \subset C_i$ is the support set.

Readmission diagnosis prediction. We consider the readmission diagnosis prediction task to further simulate and evaluate a real-world diagnostic process. In contrast to the diagnosis prediction task, the support set $S_i \subset C_{i-1}$ consists of the diagnoses of the last clinical admission C_{i-1} of

[1] https://github.com/bvanaken/clinical-outcome-prediction

the same patient from the patients' journey that consists of m admissions $P = \{A_{0_p}, \ldots, A_{m_p}\}$, where $A_i = (T_i, S_i, C_i)$. T_i is the admission note and $C_i \subset \mathbb{C}$ is the set of diagnostic codes. The motivation behind this is to integrate prior knowledge about the patient from his former admissions at the same hospital. As an additional difficulty, the model must compensate at this point for the fact that the codes from the previous admission are not necessarily supporting the diagnosis prediction, nor do the codes from the current admission functionally depend on them.

Clinical admissions and discharge summaries. We use the freely available Medical Information Mart for Intensive Care v1.4 database (MIMIC-III) (Johnson et al., 2016), containing de-identified electronic health record data (EHR), including textual discharge summaries in English of the Beth Israel Deaconess Medical Center in Massachusetts between 2001 and 2012. Following van Aken et al. (2021), we filter those textual discharge summaries by sections known at admission time, like "Chief complaint," "Medical history," or "Admission medications." The diagnostic codes associated with those admissions are using the ICD-9-CM format. Since ICD-9-CM is a very fine-grained medical coding standard, we aggregate the label space using the Clinical Classifications Software (CCS) for ICD-9-CM [2], which merges similar ICD-9-CM codes into a categorical group. Table 1 provides an overview of the dataset statistics. We also use MIMIC-III for the task of readmission diagnosis prediction but focus only on patients with more than one admission. Statistics about this subset are shown in Table 2.

	Total	Train	Val	Test
Admissions	48741	33994	4918	9829
Min. tokens	28	29	31	28
Max. tokens	17034	17034	4039	3304
∅ Tokens	641	640	635	647
Min. diagnoses	1	1	1	1
Max. diagnoses	34	34	33	33
∅ diagnoses	10.41	10.40	10.32	10.50
Unique diagnoses	280	279	266	272

Table 1: Statistics of our dataset for the task of diagnosis prediction. Very rare codes might appear only in one of the three splits.

	Total	Train	Test
Admissions	18785	13785	5000
∅ Diag. / patient	11.55	10.87	13.45
∅ New diag. / patient	8.03	7.58	9.28
∅ Lost diag. / patient	3.33	3.25	3.56
∅ Persistent diag. / patient	3.52	3.28	4.17
Unique diagnoses	270	268	256

Table 2: Statistics of our dataset for the task of readmission diagnosis prediction. On average, patients keep 3.52 of their previous diagnosis codes only. Between two admission, a patient no longer shows symptoms for an average of one-third of their previously annotated diagnosis codes.

4 Models

Augmenting text with set embeddings. For the task of set-augmented diagnosis prediction, we require a network architecture that is able to combine two possibly complementary information sources from different modalities: Clinical text from admission notes and sets of diagnosis codes. Also, the architecture must be able to learn a meaningful representation from a few examples without catastrophic forgetting in the underlying pre-trained language model. The attention mechanism (Bahdanau et al., 2015; Kim et al., 2017) allows the model to base its decision on a fine granular selection of the information in the two input spaces and to ignore less important elements. Thus, it enables the model to enrich the incomplete text representations with knowledge from the support set.

4.1 Novel Architectures

We apply three different transformer-based architectures (s. Figure 3) to incorporate knowledge from support sets to enhance the models' prediction. We preserve the permutation invariance of the set of added codes by feeding them directly into the transformer and omitting the positional encoding. Moreover, following (Devlin et al., 2019), we add the special token `[NULL]` to every support set. The `[NULL]` token also serves as an aggregate representation for the support set.

Pooled Attention. In the *pooled attention architecture* (s. Figure 3) we use the last hidden state of the `[CLS]` token as a pooling mechanism $pool()$ to aggregate the information from all tokens in the text into a single embedding. With the pooled attention architecture, we aim to compress the admission note into a meaningful single vector text representa-

[2] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

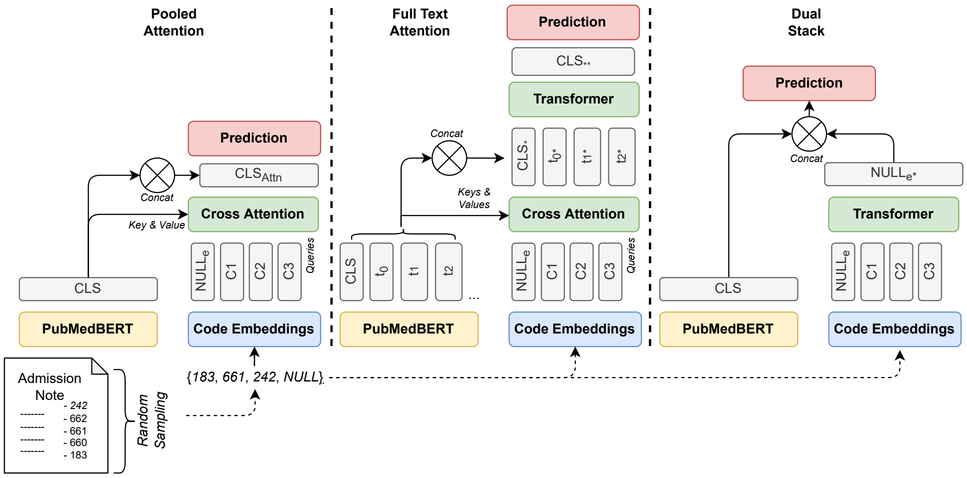


Figure 3: Support Set Augmentation Architectures with different levels of attention between text and support code features showing the Pooled Attention, Full Text Attention and Dual Stack architecture from left to right.

tion that contains all necessary information to solve the diagnosis prediction task. We project the text T_i from admission A_i into an embedding G.

$$G = pool(BERT(T_i)) \tag{1}$$

Furthermore, we apply attention between the `[CLS]` token and the input code embeddings. We use a transformer to learn shared features between the elements in the support set. Typically, the attention mechanism uses a softmax function to normalize the attention scores (Vaswani et al., 2017). However, the softmax limits the information flow to a single code embedding. Following Gülçehre et al. (2019), we replace the softmax function with a sigmoid activation σ and define our cross attention with queries Q, keys K and values V as follows:

$$\begin{gathered}\text{Cross Attention} = \sigma(QK^T)V \\ Q = G \cdot W_Q, K = G \cdot W_K, V = D \cdot W_V\end{gathered} \tag{2}$$

We denote the admission note representation by Q and K and the code representations by V and linearly transform them with learned weight matrices W_Q, W_K and W_V. The sigmoid function allows information to flow between all code embeddings and the admission note representation. Finally, we use a skipthrough connection and concatenate the output of the attention layer with the `[CLS]` representation to minimize information loss and to avoid catastrophic forgetting on the text encoder side.

Full Text Attention. With the *full text attention architecture*, we aim to reduce the potential information loss in the aggregation step in contrast to the pooled attention model. The architecture applies softmax attention between all tokens of the admission note and all codes in the support set. We use an additional transformer on top of the attended admission note tokens and use the resulting output of the `[CLS]` token for the prediction step. In distinction to the pooled attention, we only add the `[NULL]` token for empty support sets S_i.

Dual Stack. Finally, we experiment with a less complex and, compared to the full text attention model, computationally more efficient *dual stack architecture* that does not involve an attention mechanism to mix the support set with the text embedding. Instead, it consists of two independent encoders: one for the admission note and one for the support set. We use a BERT architecture as the admission note encoder and train a multi-head transformer for the support set representation. To combine the information from both information spaces, we concatenate the embeddings from both encoders and feed them into the prediction layer.

Loss function. We optimize all models by minimizing the multi-label binary cross entropy loss between the predictions p and the target labels y:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij}) \tag{3}$$

where M is the number of diagnosis codes and N the number of admission notes.

4.2 Augmenting Codes with Ontologies

The diagnosis code distribution from Figure 1 shows that there are very few training samples for codes from the long tail. This raises the concern that the model might not be able to learn a meaningful representation for each code. We initialise our diagnosis sets with additional textual information about diagnosis codes from medical ontologies to address this issue. We use three different data sources: **(1)** We obtain the descriptive name for every diagnostic code from the MIMIC-III database and map it to the respective CCS code. In addition, **(2)** we use textual definitions from the Unified Medical Language System (UMLS) 2021AB (Bodenreider, 2004), which provides comprehensive full-text definitions for 22.1% of the diagnostic codes present in our data. Furthermore **(3)**, for additional 4.7% of the codes we use descriptions from the Wikidata knowledge graph.

4.3 Baselines

We compare our approaches against two powerful baselines representing state-of-the-art approaches to diagnosis prediction with either text or set data.

Baseline: PubMedBERT We compare our approach to a PubMedBERT(Gu et al., 2021) based classifier which incorporates only textual information from the admission note N_i. We use PubMedBERT instead of ClinicalBERT (Alsentzer et al., 2019) because ClinicalBERT was pre-trained on MIMIC-III notes and therefore already contains knowledge from the discharge notes, leading to an unfair advantage in the diagnosis prediction task from admission notes only. We fine-tune the model and use the last hidden state of the [CLS] token to predict the diagnostic codes.

Baseline: Support Set Transformer This architecture only incorporates knowledge from the support set. We use the transformer based architecture from the dual stacks' support set encoder (s. Figure 3) and use only the support set S_i from the admission to predict the remaining annotated diagnostic codes C_i. Similar to the dual stack approach, we aggregate the information of the input set into a single embedding by adding a special token [NULL].

4.4 Hyperparameter Setup

We use PubMedBERT (Gu et al., 2021) as a text encoder for all text-related components of our architecture such as the admission note encoder or the ontology-knowledge augmented set encoder. We use the Adam optimizer (Kingma and Ba, 2015) with a weight decay of 0.01. Our code and hyperparameters are publicly available[3]. Furthermore, we performed a hyperparameter optimization for all architectures and also report details regarding the tuned parameters in the appendix. To prevent catastrophic forgetting in the pre-trained text encoder, we use a lower learning rate of 2e-5 for the weights of the BERT model. Due to the sequence length limitation of PubMedBERT, we truncate all admission notes to 512 tokens. We use a code embedding $\in \mathbb{R}^{768}$. The transformer in the full text attention model consists of four layers with two attention heads each. The dual stack model uses a transformer composed of one attention layer with 12 heads. We sample three annotated codes from the admission note for the diagnosis prediction task, which produced optimal results during training based on our HPO. In the readmission task we use all codes from the previous admission or the [NULL] token for the first admission.

5 Evaluation

Metrics. We measure the performance of our experiments in macro averaged *AUROC* (area under the receiver operating characteristic curve) and *mAP* (mean average precision). Because the supplied support set S_i in the diagnosis prediction task is part of the target label space $S_i \subset C_i$, it provides the support set augmented architectures an unfair advantage over the baselines to evaluate on codes $\in S_i$. Therefore, we only evaluate our approach on $y = C_i \setminus S_i$ to exclude the advantage of provided codes and to avoid that codes from the support set S_i are determined as correct predictions.

5.1 Results

We report scores of our quantitative evaluation in Table 3 and use a support set of five codes for the diagnosis prediction task to augment the admission note. *Set Embeddings* denote the combined representation of set and text data. *Semantic Set Embeddings* contain codes enriched with ontology knowledge as described in Section 4.2 and the admission notes' text. In addition to mAP and AUROC, we also report the standard error over five runs for the diagnosis prediction task because it involved random sampling to generate the support sets.

Novel models outperform baselines. We report that all of our approaches outperform the baselines,

[3] https://github.com/DATEXIS/ClinicalSupportSetAugmentation

		Diagnosis Prediction		Readmission Task	
	Model	AUROC	mAP	AUROC	mAP
Baselines	Support Set Transformer	75.52 ±1.7e-3	30.51 ±5.6e-4	66.66	44.33
	PubMedBERT	84.67 ±7.0e-4	47.39 ±7.8e-4	79.98	59.32
Set Embeddings	Full Text Attention	86.93 ±2.9e-4	49.12 ±9.2e-4	**81.37**	59.74
	Pooled Attention	87.08 ±8.2e-4	48.96 ±6.0e-4	81.06	**60.59**
	Dual Stack	87.10 ±7.0e-4	48.95 ±4.0e-4	81.01	60.54
Semantic Set Embeddings	Full Text Attention	**87.24** ±1.0e-3	**49.66** ±9.0e-4	81.03	59.61
	Pooled Attention	87.21 ±7.3e-4	48.85 ±1.5e-4	80.95	59.00
	Dual Stack	87.18 ±1.2e-3	48.67 ±2.4e-4	81.03	59.35

Table 3: Results on the diagnosis- and readmission diagnosis prediction task in macro averaged AUROC and mAP. All of our proposed architectures outperform the baselines. The full text attention model with semantic initialization of the diagnosis code embeddings performs best on the diagnosis prediction task. Semantic integration especially helps the task of diagnosis prediction, while learned set embeddings perform better for the readmission task.

emphasizing our hypothesis that augmenting admission notes with support sets containing diagnostic information helps. However, we observe that there is little difference in the performance between our proposed architectures. In general, the full text attention model performs best. Using the text and the support set in combination leads to an average improvement of around 2-3 points in AUROC or 2.5 points in mAP compared to the PubMedBERT baseline.

Minor gains with semantic set embeddings. We observe a slight increase in performance by integrating semantic knowledge (s. Table 3). However, the co-occurrence of codes within the admissions seems to have a much more substantial impact on the final classification performance than the additional semantic information. This slight increase indicates that the architecture can leverage the additional information, but the semantics encoded in the ICD names and UMLS definitions do not seem to contain much complementary knowledge.

Rare and very frequent codes are most effective. We analyze the impact of the diagnosis code frequency on the prediction performance. We perform ten evaluations with each three random sampled codes in the support set, binned by frequency. We measure the performance difference between our model and the PubMedBERT baseline on the remaining codes and plot the standard error for those observations (s. Figure 4). We observe that both rare codes from the end of the long-tail and frequent codes belonging to the second tertile improve the prediction by almost five points in mAP. We hypothesize that especially rare codes create a major distinctive factor for a diagnosis where the machine assigns a high weight. Contrary, obvious and frequent diagnoses may create a bias and thus have the highest impact on the prediction. The smaller increase in performance between frequency bins 50 to 250 indicates partial to non-existent complementarity between text and support set. In general, the effect of diminishing performance difference with decreasing frequency can be explained by few training examples for the given code and thus an insufficiently learned representation because MIMIC-III primarily focuses on severely ill patients that require life-saving measures at the ICU.

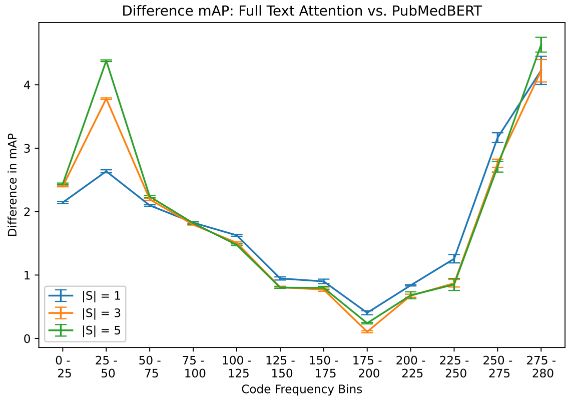


Figure 4: Performance difference between PubMedBERT and full text attention with support sets of size 1, 3, and 5 codes across different frequency bins.

Random sampling for compensating imbalance. The CCS label distribution follows a power-law distribution pattern (s. Figure 1). To compensate effects of such an imbalanced label distribution, we evaluate random sampling vs. inverse frequency weighted random sampling to create potentially more balanced support sets during training. We find that inverse frequency weighted random sampling, in general, performs worse than random sampling

$\|S_i\|$	**0-13**	**13-43**	**43-280**
0	80.66	82.03	86.41
1	81.61	83.19	87.94
2	82.08	83.61	**88.36**
3	**82.54**	**84.03**	87.91
PubMedBERT	79.28	81.09	85.73

Table 4: Prediction performance over different code frequencies measured in AUROC with different support set sizes using random sampling and the full-text attention model with set embeddings. Frequencies split in tertiles (s. Figure 1) according to the code distribution in the admission note dataset. The performance increases with growing support set sizes $|S_i|$.

that follows the label distribution in the training data. We observe a difference in mAP of *-0.3*, *-1.8*, and *-1.3* points for the full text, pooled, and dual stack architectures compared to training with random sampling. Our explanation is that random sampling focuses more on the short head of the label distribution. Therefore, the model learns a richer representation for codes that are more frequent in the dataset and therefore provides better support for common predictions (s. Figure 4).

Larger sets can be beneficial. Table 4 shows that increasing the number of elements in the support set improves the performance for codes of all frequencies. Especially in the range of the 25-50 most frequent codes (s. Figure 4), larger support set sizes lead to the most performance improvement. Finally, we observe that even with $|S_i| = 0$, our model outperforms the PubMedBERT baseline, which indicates that the model stores valuable information in the `[NULL]` token.

6 Discussion

Clever Hans problem for readmissions. Language models often just learn effective shortcuts of high dimensional data distributions instead of generalizing, which is called *Clever Hans problem* (Lapuschkin et al., 2019). We expect to find a variant of the Clever Hans problem for the readmission diagnosis prediction task: Here, we expect the model to learn the shortcut of copying codes from the support set into the diagnosis predictions instead of learning novel correlations from the potentially complementary text data. Indeed, our model copies in 78.1% of the test cases, on average, 2.12 codes from the support set that are not in the target label set. However, this is only a tiny fraction of the average of 13.45 codes in each support set (s. Table 2). This contradicts the Clever Hans problem and empirically confirms the ability of our model to ignore unrelated information from the support set and, in those cases, to focus more on the admission note.

Beneficial and non-beneficial codes. In our evaluation with clinical doctors we observe certain codes that improve the prediction more than others. We find that these codes are typically rare, such as code *188* (s. Table 5), but improve the mAP relative to PubMedBERT by more than 30 points. Likewise, rare codes can also have a diminishing effect on the prediction performance. We hypothesize that their representation is not well initialized due to the lack of training examples. In Table 5, we show the most helpful and most unhelpful codes and their rank in the dataset. We find that 221 codes improve the prediction by, on average, 3.02%. 51 codes decrease the prediction performance by, on average, 2.29%.

CCS Code	**ΔmAP**	**Rank**
188 - Fetopelvic disproportion; obstruction	+32.9	269
187 - Malposition; malpresentation	+25.8	259
191 - Polyhydramnios and other problems of amniotic cavity	+24.0	270
31 - Cancer of other male genital organs	+23.4	272
184 - Early or threatened labor	+21.3	242
..		
119 - Varicose veins of lower extremity	-05.2	234
218 - Liveborn	-05.4	267
655 - Disorders usually diagnosed in infancy, childhood, or adolescence	-05.9	247
124 - Acute and chronic tonsillitis	-09.6	257
177 - Spontaneous abortion	-15.3	261

Table 5: Excerpt of all codes in the support set that have the most effect on prediction performance compared to PubMedBERT ranked by mAP difference. Codes from the long tail have the highest impact on performance.

Additional set data can compensate problems of language models with idiosyncratic language. Commonly, large language models often have difficulties with domain-specific language (Liu et al., 2020). For example, pregnancy is often encoded in an idiosyncratic manner like "G2P1," which stands for *gravida 2 para 1*, which means that this is the second pregnancy and the first one's result was a life-born child. Also, abbreviations such as *"PNV"* for prenatal vitamins are usually the only indicator of pregnancy. Often language models do not have seen sufficient context information during training to generalize from these words to infer higher-level concepts. We analyze the effect of the most influential codes, from which 15 of 20 are either pregnancy or gynecology related. We measure the difference in rank with and without these codes in the support set. Of all the codes that benefit from adding these codes, 73.90% are pregnancy or gynecology-related as well. Our results indicate that adding pregnancy-related codes to the support set helps the model recognize the concept of pregnancy and compensates for problems of idiosyncratic language.

Future improvements. It is interesting to see if graph neural networks can lead to improved representation through updates of related codes. Furthermore, it is possible to experiment with negative examples of diagnostic codes and additional encodings to represent the time between two or more admissions. Given the sparse training data situation presented in medical data silos, particular focus should be applied to zero-shot or few-shot cases, e.g., codes that occur the first time or are rarely represented in prior admissions.

7 Conclusion

Augmenting text with set data is an important problem, in particular in the clinical domain with multimodal patient representations. To solve this problem, we propose novel attention-based network architectures. Our results clearly show that in a clinical prediction task, the augmented representation outperforms a language model, particularly for predicting less common diseases. We also observe that complementary data from sets can for compensate shortcomings of language models, such as idiosyncratic language or abbreviations.

8 Ethical Considerations

Models for diagnosis prediction based on clinical admission notes can be a valuable component of clinical decision support systems that aim to assist medical professionals during their differential diagnosis. Hence, those models bear the potential to save lives by preventing inexperienced doctors from overlooking rare or unusual symptoms. They might as well save cost and reduce the amount of time required for the diagnosis process of medical professionals. However, admission notes and billing codes such as ICD-9 are only a very limited and biased perspective on the patient. Admission notes leave out important diagnostics performed during the patients' stay. Furthermore, billing codes are a suboptimal target label space. They are used to obtain the maximum possible reimbursement for the cost of treatment. There is a risk that patients will receive an excessive number of codes and, therefore, might be over-coded. Likewise, under-coded patients may also occur. To deduce clinical outcome solely from admission notes without having medical professionals perform an iterative differential diagnosis process raises the concern that certain very significant signals may never be introduced to the model.

9 Acknowledgements

We would like to thank the editors and anonymous reviewers for their helpful suggestions and comments. Furthermore, we would like to thank Tom Röhr for his support throughout the project. This work was founded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreements 01MK2008MD (Service-Meister) and 01MD19003B (PLASS), as well as the Federal Ministry of Education and Research (BMBF) under grant agreement 16SV8845 (KIP-SDM).

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. *CoRR*, abs/1904.03323.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Linkun Cai, Yu Song, Tao Liu, and Kunli Zhang. 2020. A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. *IEEE Access*, 8:152183–152192.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3504–3512.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric J. Topol, Jeff Dean, and Richard Socher. 2021. Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4.

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

Çaglar Gülçehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter W. Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. 2019. Hyperbolic attention networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Alistair EW 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):1–9.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. 2016. Learning to diagnose with LSTM recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep EHR: chronic disease prediction using medical notes. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2018, 17-18 August 2018, Palo Alto, California*, volume 85 of *Proceedings of Machine Learning Research*, pages 440–464. PMLR.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1903–1911. ACM.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 743–752. ACM.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10.

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 Language Resources and Evaluation Conference*, pages 900–909, Marseille, France. European Language Resources Association.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. 2020. Self-attention enhanced patient journey understanding in healthcare system. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, volume 12459 of *Lecture Notes in Computer Science*, pages 719–735. Springer.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. MNN: multimodal attentional neural networks for diagnosis prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5937–5943. ijcai.org.

Madhumita Sushil, Simon Suster, Kim Luyckx, and Walter Daelemans. 2018a. Patient representation learning and interpretable evaluation using clinical notes. *J. Biomed. Informatics*, 84:103–113.

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

Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Bo Yang and Lijun Wu. 2021. How to leverage the multimodal EHR data for better medical prediction? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4029–4038. Association for Computational Linguistics.