EXBEHRT: EXTENDED TRANSFORMER FOR ELECTRONIC HEALTH RECORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this study, we introduce ExBEHRT, an extended version of BEHRT (BERT applied to Electronic Health Record data) and applied various algorithms to interpret its results. While BEHRT solely incorporates diagnoses and age of patients, we extend the feature space to several multi-modal records, namely demographics, clinical characteristics, vital signs, smoking status, diagnoses, procedures, medications and lab tests using a novel method to unify the frequencies and temporal dimensions of the different features. We show that additional features yield significant benefit in model performance for various down-stream tasks of different diseases. To ensure robustness, we interpret the model predictions using an adaption of Expected Gradients, which hasn't been applied to Transformers with EHR data so far and yields more granular interpretations than previous approaches such as feature and token importances. Further, by clustering the models' representations of oncology patients, we show that the model has implicit understanding of the disease and is able to divide patients suffering from the same cancer type into different risk groups. Given the additional features and interpretability, ExBEHRT can help making informed decisions about disease progressions, diagnoses and risk factors of various diseases.

1 INTRODUCTION

Over the last decade, Electronic Health Records have become increasingly popular to document a patients' treatments, labs, vital signs, etc. Commonly, a sequence of medical events is referred to as a *patient journey*. Given the immense amount of longitudinal data available, there lies tremendous potential for Machine Learning to generate novel insights about the recognition of disease patterns, progression and subgroups as well as treatment planning. Recent studies adapted Transformers to structured EHR data, demonstrating superiority in various benchmarks compared to other, similar algorithms. The first adaptation of Transformers to EHR data, called BEHRT (Li et al., 2020), incorporated diagnosis concepts and age from EHRs and additionally added embeddings for the separation of individual visits and a positional embedding for the visit number. Other models such as Med-BERT (Rasmy et al., 2021), CEHR-BERT (Pang et al., 2021) and BRLTM (Meng et al., 2021) added more features by concatenating the inputs into one long patient sequence. These approaches are limited in the amount of data of a single patient they can process and the required computational power significantly increases with each added feature.
In this work, we introduce a novel approach of incorporating multi-modal features into Transformer models by adding medical concepts separately and vertically instead of concatenating all concepts horizontally. We show that these features are important in several downstream applications such as mortality prediction, patient subtyping and disease progression prediction.

2 EXBEHRT FOR EHR REPRESENTATION LEARNING

ExBEHRT is an extension of BEHRT where medical concepts are not concatenated into one long vector, but rather grouped into separate, learnable embeddings per type of concept. This way, we avoid exploding input lengths when adding new medical features and provide the model the capability to learn which concepts it should focus on. Clinically, it would also be stringent to separate diagnoses, procedures, medications etc. as they offer different clinical value for downstream applications. We take the number of diagnoses in a visit as the indicator of how many "horizontal slots"

for other concepts are available at this visit (e.g. two for the first visit in figure 1). Therefore, the maximal patient journey length is defined by the amount of diagnosis codes of a patient, independent of the amount of other concepts added to the model. Another advantage of this procedure is its ability to deal with varying frequencies and sparseness of additional concepts. As exemplified with procedures in figure 1, but executed in the same manner with labs, there are three possible cases of adding a new concept to a visit:

1. The number of procedures is equal to the amount of horizontal slots available in the visit (visit 1 - two each). The procedures can therefore be represented as a 1D vector.
2. The number of procedures exceeds the amount of slots available in the visit (visit 2 - one diagnosis, two procedures). Here, the procedures fill up the amount of horizontal slots in a row-wise manner until there are no procedures left, resulting in a 2D vector of dimensions $\#slots \times \lceil \frac{\#procedures}{\#slots} \rceil$.
3. The number of procedures subceeds the amount of slots available (visit 3 - one diagnosis, no procedures). Procedures are represented as a 1D vector and then padded to the amount of horizontal slots available.

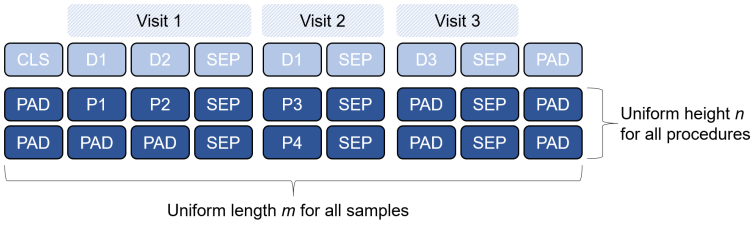

Figure 1: An example of how ExBEHRT represents a patient with a constant sentence length m.

After reshaping, all procedures and labs of all patients are padded to the same amount of rows n to enable batch processing. Before passing the inputs to the model, each token is embedded into a 288-dimensional vector and all tokens are summed vertically. Figure 7 in the appendix shows the final representation of one patient.

2.1 Data

In this study, we used one of the largest EHR datasets from the USA. We only selected data points collected at hospitalizations to ensure data quality and consistency. Each patient is required to have at least five visits with valid ICD-9 or ICD-10 diagnosis codes to ensure sufficient temporal context. Given these criteria, our final pre-training cohort consisted of 5.4M individual patients split into training (80%), validation (10%) and testing sets (10%).

2.2 Model Training

ExBEHRT consists of the same model architecture as BEHRT. For pre-training, we applied the standard MLM procedure described in the original BERT paper (Devlin et al., 2018). In a second step, we fine-tuned our model on two prediction tasks: Death of a patient within six months after the first cancer diagnosis and readmission into hospital within 30 or less days after a heart failure. All tokens after the cancer diagnosis/heart failure are not disclosed to the model. We further used the patient representations of ExBEHRT to identify risk subtypes of cancer patients using unsupervised clustering. For that, we applied a combination of the dimensionality reduction technique UMAP (McInnes et al., 2018) and the clustering algorithm HDBSCAN (Campello et al., 2013).

3 Results

3.1 Event Prediction

In all but one metric in one task, ExBEHRT outperforms BEHRT and other conventional algorithms such as Logistic Regression (LR) and XGBoost when evaluated on the hold-out test set.

Table 1: Test set micro-averaged metrics for fine-tuning on event prediction tasks.

Task	Cohort Size	Metric	LR	XGBoost	BEHRT	ExBEHRT
Death in 6M	Train: 350'322	APS	0.4280	0.4554	0.4778	**0.5362**
	Val: 43'790	AUROC	0.6345	0.6642	0.6697	**0.7255**
	Test: 43'790	Precision	0.7304	0.7431	0.7520	**0.7824**
HF readmit	Train: 402'529	APS	0.2976	**0.3132**	0.1995	0.2501
	Val: 50'316	AUROC	0.5190	0.5359	0.5117	**0.5670**
	Test: 50'317	Precision	0.7199	0.7258	0.8102	**0.8163**

3.2 Interpretability on Event Prediction Results

For all interpretability experiments, we used our model fine-tuned on the task *Death in 6M*, meaning whether a cancer patient will decease within six months after their first cancer diagnosis. We visualize the interpretability for single patients as both interpretability approaches are example-based and not model-agnostic.

3.2.1 Self-Attention Visualization

Analogously to previous papers (Li et al. (2020), Rasmy et al. (2021), Meng et al. (2021)), we visualized the attention of the last network layer using BertViz (Vig (2019)). However, since all embeddings are summed before being passed through the network for all such models, self-attention has no possibility to attribute single input features to the outcome. Nevertheless, we can deduce insights on how the different slots interact with each other and which connections the model deems to be important. Figure 2 displays the self-attention of a single patient of the last layer of ExBEHRT. The left figure displays the attentions of all 12 attention heads in this layer, whereas the right figure displays the attention of one head. Generally, the model focuses a lot on the slots within one visit, which was to be expected as these slots are strongly related by definition. Slot 7 corresponds to the slot, where the patient was diagnosed with lung cancer. Even though the model was not specifically trained to put emphasis on cancer codes, it puts a great amount of attention on this slot, indicating that it learned some correlation between the cancer diagnosis and the predicted outcome. Interestingly, slot 7 receives a lot attention from the first and the second visit, but not from the other two previous visits, indicating that the model is able to learn causalities across long time gaps.

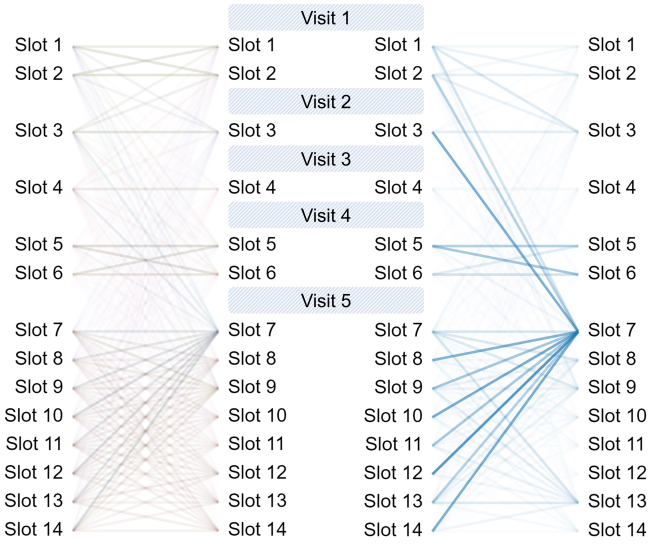

Figure 2: Left: The self-attention of all 12 attention heads of the last layer of ExBEHRT. Higher opacity corresponds to higher attention. Right: The self-attention of one attention head of the last layer. Slot 7 corresponds to the slot where the cancer was diagnosed.

3.2.2 EXPECTED GRADIENTS INTERPRETABILITY

Due to the limitations of Self-Attention visualization, we explored the technique Expected Gradients (Erion et al., 2020) for deeper understanding of the model. Expected Gradients is considered to be one of the most robust gradient-based feature attribution methods for deep learning models. This way, we can deduce feature and token importances of single predictions, which is not possible with self-attention. Since each single concept (diagnosis code, procedure code, age, etc.) is mapped to a 288-dimensional embedding before being passed to the model, we first calculated the expected gradients for each of the 288 positions and then summed the absolute values to acquire a single gradient value for each input token. This way, each individual input token has an associated gradient related to the models output yielding detailed insights into which medical concept has had what impact on the models prediction. We visualized the findings in three layers of abstraction with increasing detail in figures 3, 4 and 5.

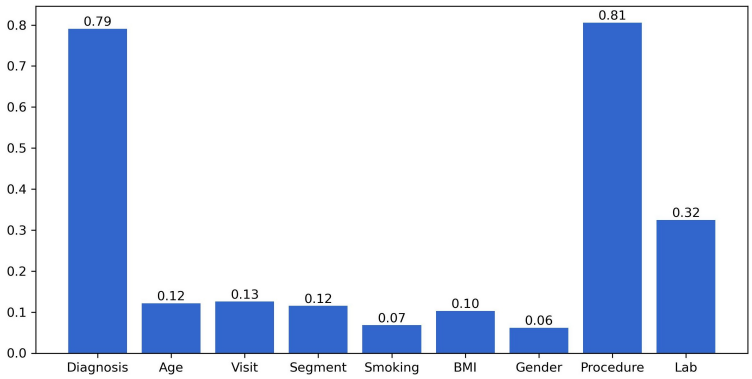

Figure 3: The absolute sums of the expected gradients summed by input feature.

For figure 3, we summed the expected gradients for each of the input features. This way, we can evaluate the different feature impacts on the output for a specific patient. For this patient, diagnoses and procedures (treatments & medications) were by far the most importance features. With this visualization, we can further evaluate basic biases. For example, gender was not considered to be an important feature, indicating that performance would be similar for a person with another gender.

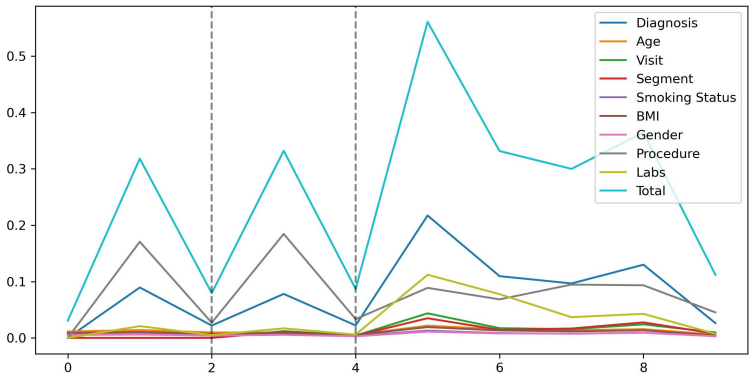

Figure 4: The absolute sums of the expected gradients summed by input feature and time slot. The dotted lines represent slots with SEP tokens and therefore indicate the next visit.

For figure 4, we visualized the absolute expected gradients for each of the input features and summed them at each time slot. This way, we can evaluate the different feature importances over time to get a notion of where the model puts emphasis on. Interestingly, the model put more importance on what kind of medications & treatments that patient received in the first two visits, where as in the last visit (the visit in which the patient was diagnosed with blood cancer), it put more importance on diagnoses and labs. Generally, slot 5, where the cancer was diagnosed, was attributed with the highest importance.

Diagnosis									
CLS	J40	SEP	M54	SEP	C81	R55	R59	E87	SEP
Lab									
-	-	-	-	-	CHEMISTRY	URINALYSIS	HEMATOLOGY	SPEC. CHEM.	-
-	-	-	-	-	SPEC. LAB	BLOOD GAS	-	-	-
Procedure									
-	71020	-	81003	-	-	-	-	-	-
-	94640	-	87077	-	-	-	-	-	-
-	99283	-	87086	-	-	-	-	-	-

Figure 5: A visualization of the absolute sums of the expected gradients of diagnoses, labs and procedures on a concept level. Darker colours represent higher values and the SEP tokens indicate the separation between two visits.

Figure 5 displays the absolute sums of gradients of each individual input token, providing a detailed interpretation of which medical concept has had what impact on the models prediction. Unsurprisingly, the cancer code C81 has had the biggest impact on the outcome. However, earlier codes like J40 or 71020 also contribute to the models prediction, indicating that the model is able to include information from the whole patient journey into its results.

3.3 Cancer Patient Clustering

HDBSCAN was able to cluster 90% of all cancer patients into 24 clusters[1]. On average, the most occurring cancer diagnosis within a cluster was present for 84% of the patients assigned to this cluster and the mean cluster purity[2] was 85%. Similar concepts (e.g. cancer of female reproductive organs or different types of leukaemia) lay in areas close to each other, indicating a spatial logic between the cancer types.
In figure 6, we show that with a second pass of HDBSCAN on a specific cluster, we can identify risk subgroups. In all three identified clusters, more than 90% of the patients actually do have pancreatic cancer and all clusters share similar general characteristics. However, as displayed in table 3.3, ExBEHRT identified one subgroup with significantly higher chance of recovering from cancer and a lower probability of death, even though this information was not provided to the model at any point[3].

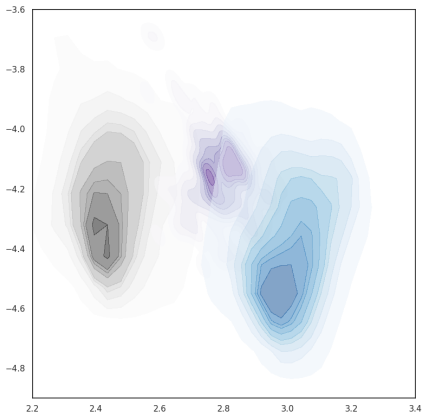

Figure 6: The three identified patient subclusters with pancreatic cancer visualized with a kernel density estimate plot for visual clarity.

[1]A visualization of all clusters can be found in figure 8 in the appendix.

[2]Cluster purity indicates the fraction of patients with a condition which are assigned to the cluster.

[3]In the table, *% of journey with cancer* indicates the ratio of the time between the first and last cancer diagnosis compared to the duration of the whole recorded patient journey. *Cancer-free* refers to the percentage of patients within a cluster, which have records of at least two visits without cancer diagnosis after the last visit with a cancer diagnosis. The *average death rate* is directly taken from the EHR database and unfortunately does not indicate the cause of death.

Table 2: Statistics of the three pancreatic cancer clusters indicating a clear differentiation between higher risk (gray, blue) and lower risk patients (purple).

Metric	**Gray**	**Blue**	**Purple**
Median age	67	68	68
Median birth year	1950	1947	1944
Median BMI	25	25	26
Average death rate	76.5%	75.9%	**70.0%**
% of journey with cancer	27.0%	24.0%	**18.3%**
Cancer-free	34.0%	36.9%	**62.7%**

4 Discussion

In this study, we introduced a novel method of adding patient features to BEHRT, which significantly boosts the predictive power for several downstream tasks in multiple disease areas. The novel method of stacking features vertically yielded improvements in hardware requirements and benchmarks and eases the possible extension to new concepts in the future. Given the vast amount and heterogeneity of patients the model was pre-trained with, we are confident that ExBEHRT would generalize well to novel data, patients and tasks. Combined with interpretability, the model provides more granular insights into disease progressions and subtypes of various patients than previous approaches, which could help clinicians forming more granular assessments of the progression and health of their patients. In addition, it is possible to detect unmet needs and improve patient outcomes with a personalized understanding of patient groups.
Nevertheless, there are a few limitations: It is extremely difficult to validate the quality, completeness and correctness of EHR datasets, as the data is usually processed anonymously and stems from a variety of heterogeneous, fragmented sources. There is also bias in the sheer nature of EHR data, as practitioners could be incentivised to additionally diagnose less relevant conditions, since medical billing is closely related to the amount and type of diagnoses indicated. In a potential next step, we would like to verify the results and interpretations of this work with clinicians to ensure robust and sound predictions as possible given the acquired interpretability.

References

Ricardo J G B Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, Su-In Lee, and Paul G Allen. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. 2020. URL https://github.com/suinleelab/attributionpriors.

Yikuan Li, Shishir Rao, Jos Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. 12 2020. ISSN 20452322. doi: 10.1038/s41598-020-62922-y.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. 2 2018.

Yiwen Meng, William Speier, Michael K. Ong, and Corey W. Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. pp. 3121–3129, 8 2021. ISSN 21682208. doi: 10.1109/JBHI.2021.3063721.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. 11 2021.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. 2021.

Jesse Vig. A multiscale visualization of attention in the transformer model. pp. 37–42, 2019. doi: 10.18653/V1/P19-3007.

A APPENDIX

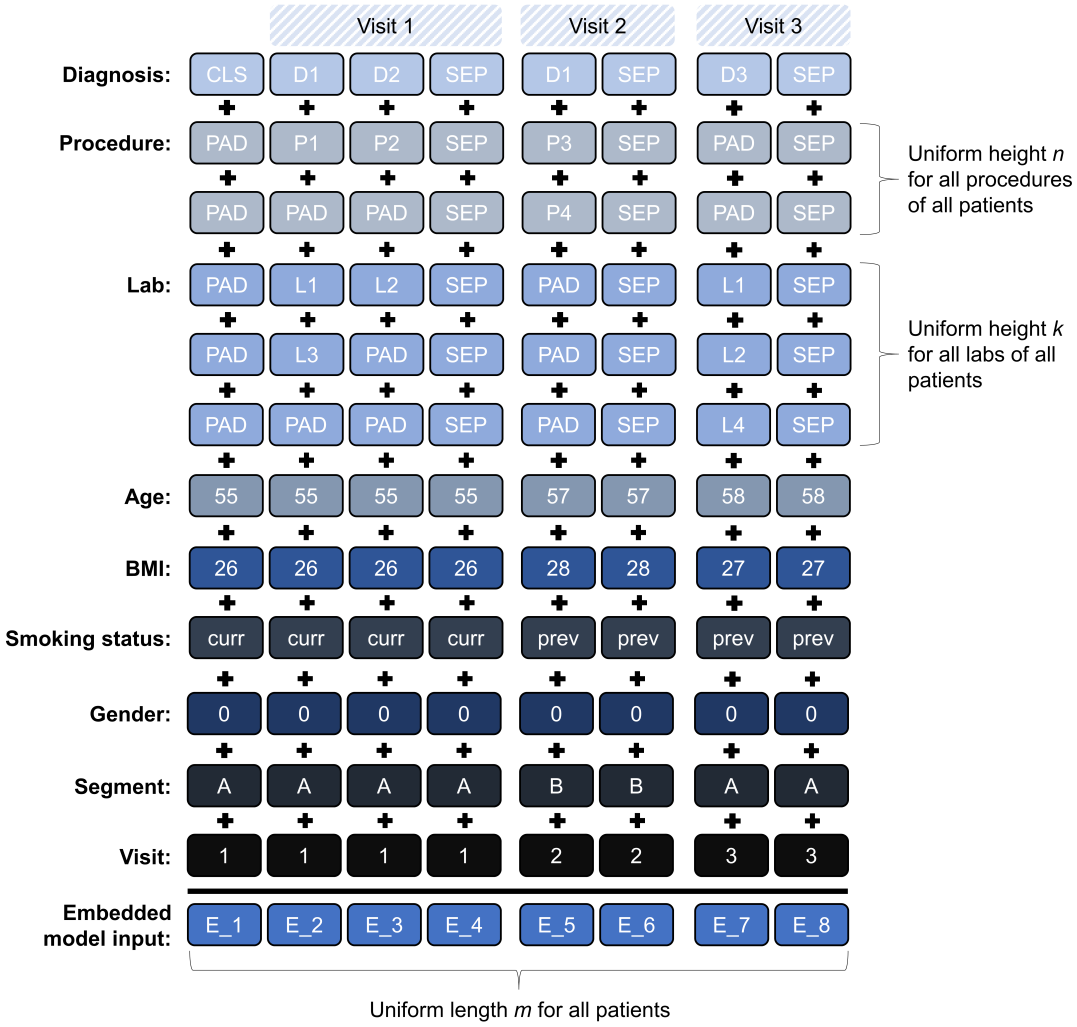

Figure 7: A sample input of ExBEHRT. Each of the concepts (diagnosis, procedure, lab, age, BMI, smoking status, gender, segment, visit) has its own embedding, where each of the tokens is mapped to a 288-dimensional vector, which is learned during model training. After embedding, all concepts are summed vertically element-wise to create a single $288 \times m$ dimensional vector as input for the model. The parameters n and k are set before training and fixed for each patient to ensure coherent batch processing.

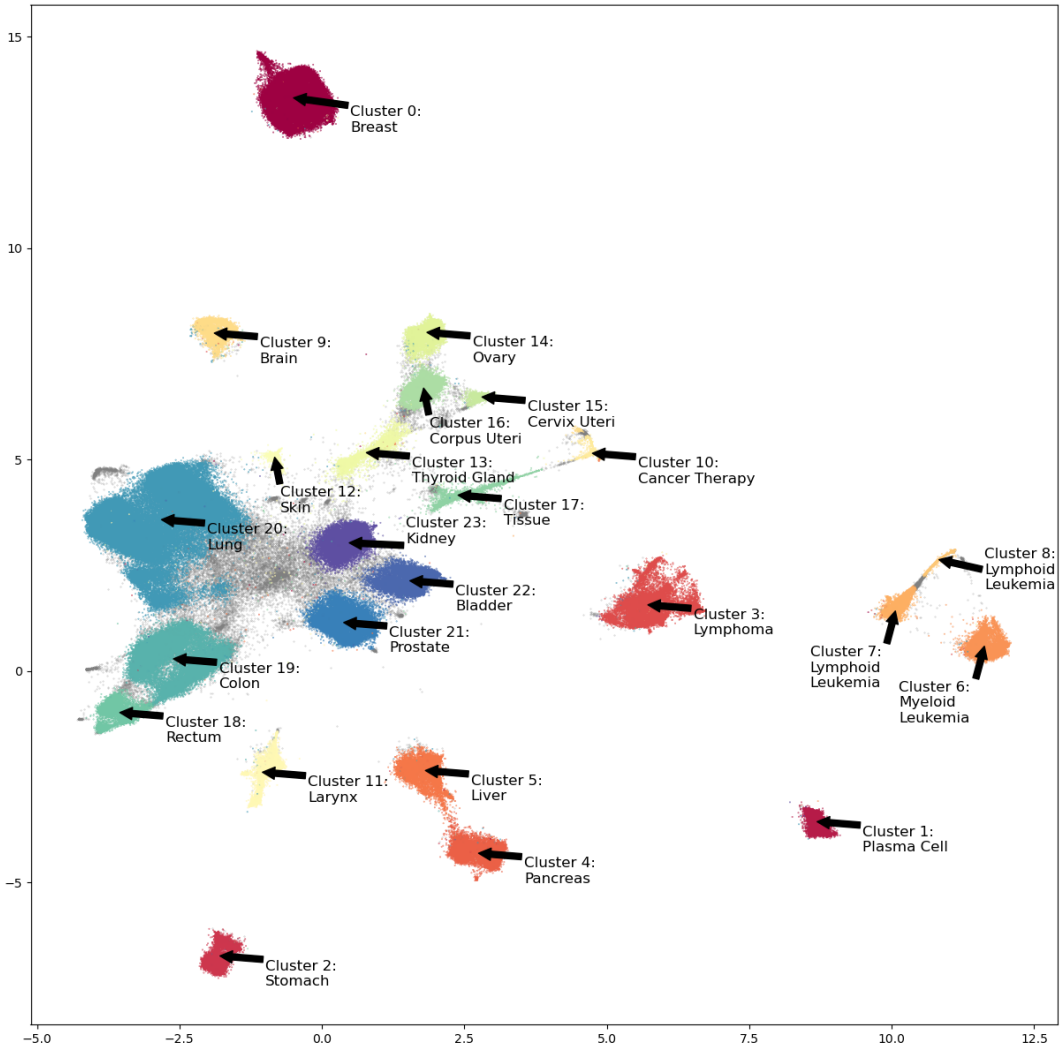

Figure 8: The unsupervised cluster assignments from HDBSCAN, visualized with a 2-dimensional UMAP projection. The grey points are patients not assigned to any cluster (10%). The labels indicate the most frequent diagnosis code of each cluster. Besides cluster 10, all label are neoplasms. On average, the most occurring cancer diagnosis within a cluster was present for 84% of the patients assigned to this cluster. The clusters are clearly separated spatially, indicating a distinct separation of the different cancer types and their representations within the model.