sEHR-CE: Language modelling of structured EHR data for efficient and generalizable patient cohort expansion

Anna Munoz-Farre
BenevolentAI
anna.munoz.farre@benevolent.ai

Harry Rose
BenevolentAI
harry.rose@benevolent.ai

Sera Aylin Cakiroglu
BenevolentAI
aylin.cakiroglu@benevolent.ai

Abstract

Electronic health records (EHR) offer unprecedented opportunities for in-depth clinical phenotyping and prediction of clinical outcomes. Combining multiple data sources is crucial to generate a complete picture of disease prevalence, incidence and trajectories. The standard approach to combining clinical data involves collating clinical terms across different terminology systems using curated maps, which are often inaccurate and/or incomplete. Here, we propose sEHR-CE, a novel framework based on transformers to enable integrated phenotyping and analyses of heterogeneous clinical datasets without relying on these mappings. We unify clinical terminologies using textual descriptors of concepts, and represent individuals' EHR as sections of text. We then fine-tune pre-trained language models to predict disease phenotypes more accurately than non-text and single terminology approaches. We validate our approach using primary and secondary care data from the UK Biobank, a large-scale research study. Finally, we illustrate in a type 2 diabetes use case how sEHR-CE identifies individuals without diagnosis that share clinical characteristics with patients.

1 Introduction

Electronic health records (EHR) are collected as part of routine medical care, and include demographic information, disease diagnoses, laboratory results, medication prescriptions, etc., providing a patient's clinical state over time. Recent machine learning techniques have been used to exploit the richness of EHR data at scale for diagnosis, prognosis, treatment and understanding of disease [Li et al., 2020, Steinberg et al., 2021]. Many medical terminologies are used across clinical data sets, and the standard practice involves mapping clinical terms across different resources [Li et al., 2020, Hassaine et al., 2020] or onto common data models [Stang et al., 2010].

Here, we propose sEHR-CE (language modelling of **S**tructured **EHR** data for patient **C**ohort **E**xpansion), a novel framework based on transformers to enable the integrated analysis of multiple clinical resources without relying on any manual curation and mapping. Using text descriptions of concepts as input, our method generalises across data modalities and terminologies, i.e. text and structured EHR. This enables us to leverage a plethora of pre-trained language models like PubMedBERT [Gu et al., 2022] to encode each patient's medical record. We ask the model to learn representations of clinical histories from diagnosed patients (cases) to predict phenotypes (such as

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

diseases). In the absence of a directly comparable model, we evaluate the performance of our model against that of Li et al. [2020], as the state-of-the-art approach for learning clinical term embeddings for future disease prediction. See Appendix A for a more complete survey.

In our validation experiments, we will show that the model can then identify individuals with missing diagnoses (controls) that share a similar clinical history with cases, indicating they might have the disease or be at risk of developing it. The contributions of our paper are **(i)** the presentation of a cohort expansion method that provides phenotype predictions outperforming non-text and single terminology approaches, and **(ii)** an in-depth qualitative evaluation demonstrating that positively predicted controls share similar clinical representations as cases, providing a high degree of evidence that these may be previously un- or misdiagnosed individuals. Finally, we demonstrate our method's potential for patient stratification by disease severity.

2 Methods

Input Generation. For each EHR source and associated ontology $a \in \mathcal{A}$, we denote the set of concepts (e.g. clinical terms) as Θ_a, and the set of text descriptions as Ξ_a. The total vocabulary of concepts and text descriptions across all ontologies is denoted by $\Theta = \bigcup_{a \in \mathcal{A}} \Theta_a$ and $\Xi = \bigcup_{a \in \mathcal{A}} \Xi_a$, respectively. For each patient, we define their full clinical history through time and across sources by the concatenation of sequences of clinical descriptions $(\xi_{\theta_1}, \ldots, \xi_{\theta_t})$, $\xi_{\theta_i} \in \Xi$, $i = 1, \ldots, t$, ordered in time. More details and an example can be found in Appendix B.1, and Figure B.1. To form the input, we process the raw text sequence of descriptions into tokens (e.g. words and sub-words) $X = W(\xi_{\theta_1}, \ldots, \xi_{\theta_t}) = (x_1, \ldots, x_n)$ under a fixed size vocabulary V with the WordPiece tokenizer W [Wu et al., 2016].

Label Generation. Let $\Delta = (d_1, \ldots, d_D)$ denote an ordered set of clinical outcomes or events, in our case disease phenotypes d_i. For each phenotype $d_i \in \Delta$, we let $\mathbf{1}_{d_i}$ be an indicator function that assigns a binary label to individuals according to the presence or absence of d_i. To define $\mathbf{1}_{d_i}$, we rely on external oracles or phenotype definitions, such as CALIBER (Appendix B.2). Figure 1 shows a schematic of input and label generation for a given patient.

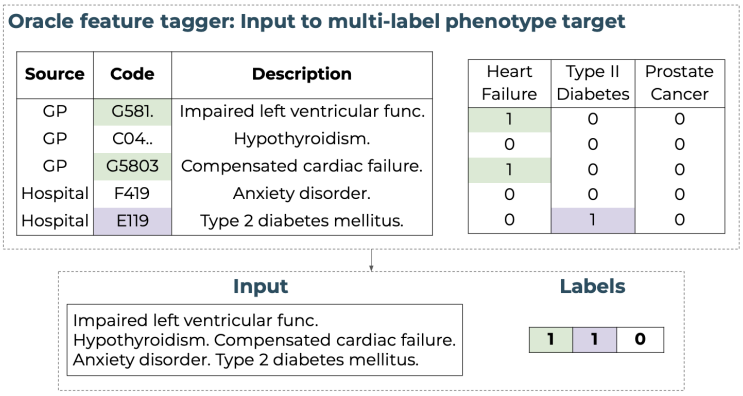


Figure 1: Schematic of mapping a unified clinical history to an input text sequence with multi-label target vector using an oracle annotation [1].

Model Design. Let $X^{(p)} = (x_1^{(p)}, \ldots, x_n^{(p)})$ denote the tokenized input sequence of individual p. It forms the input to an encoding function $\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_n^{(p)} = Encoder(X^{(p)})$, where each $\mathbf{x}_i$ is a fixed-length vector representation of each input token x_i. Let $\mathbf{y}^{(p)} = (y_1^{(p)}, \ldots, y_D^{(p)})$, $y_i^{(p)} \in \{0, 1\}$, be labels denoting presence or absence of phenotypes $d_1, \ldots, d_D$. Given a learned representation over inputs, we decode over $\mathbf{y}^{(p)}$ under the predictive model $\mathbf{P}(\mathbf{y}^{(p)}|X^{(p)})$. We calculate the probability of each phenotype d_i given the input sequence encoding $\mathbf{P}(y_i^{(p)}|\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_n^{(p)})$ via a decoder module. Specifically, we decode the representation into logits per phenotype $z_1^{(p)}, \ldots, z_D^{(p)} = Decoder(\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_n^{(p)})$ and calculate the probability per phenotype as $\mathbf{P}(y_d^{(p)}|\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_n^{(p)}) =$

$\sigma(z_d^{(p)})$, where σ denotes the sigmoid function (Figure B.2). We will omit the superscript (p) denoting the sample index in the remainder of the text.

2.1 Data Augmentation

Clinical Masking. We mask input descriptions ξ_θ from term-description pairs (θ, ξ_θ) with $\mathbf{1}_d(\theta, \xi_\theta) = 1$ for $d \in \Delta$ using the following clinical masking strategy during training and validation as in Devlin et al. [2019], Wei and Zou [2019]: **Remove** ξ_θ with 80% probability, **retain** ξ_θ with 10% probability, and **replace** ξ_θ with a randomly selected description from the corpus with 10% probability. Figure 2 displays a worked example. During testing, these term-description pairs are fully removed from the input sequence.

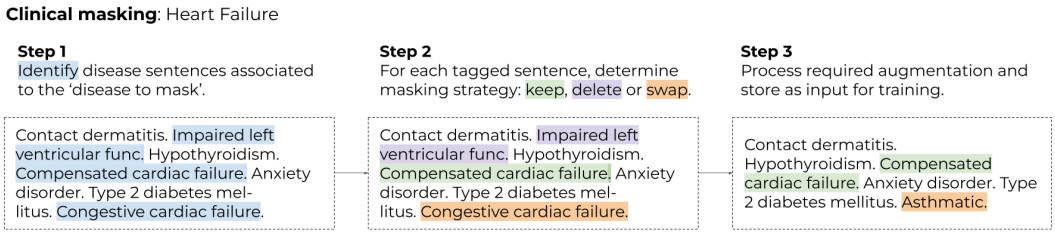


Figure 2: Example of clinical masking strategy for a given patient history with **Heart Failure**.

Comorbidities. This masking approach is straightforward if an individual has only one positive label, but many people have comorbidities, e.g. co-occurring conditions. To allow for a sample with multiple positive labels $d_{i_1}, \ldots, d_{i_n}$, we create n input samples by replicating both the input sequence and target vector of phenotype labels. Consider the jth replicated input sequence, and let $d = d_{i_j}$. We mask this replicated input sequence with the masking strategy for phenotype d described in the previous section; e.g. ξ_θ is masked if $\mathbf{1}_d(\theta, \xi_\theta) = 1$ (Figure 3).

2.2 Defining a Loss Function in the Presence of Comorbidities and varying Prevalences

Loss weights. The data augmentation approach for comorbidities may lead to our model overfitting when an input sequence contains the descriptions associated with an existing phenotype d that is not masked. To account for the contribution of the target vector $\mathbf{y}$ to the loss function in these scenario, we define a binary masking vector $\gamma^j = (\gamma_1^j, \ldots, \gamma_D^j)$ where $\gamma_j^j = 1$ and $\gamma_k^j = 0$, $k = 1, \ldots, D$, $k \neq j$. Individuals with no positive phenotype labels are assigned an all-zero masking vector (Figure 3). Then, for a given input text sequence X, target label vector $\mathbf{y} = (y_1, \ldots, y_D)$ and masking vector $\gamma = (\gamma_1, \ldots, \gamma_D)$, we define the loss weights by setting $\omega_d = 0$ if $y_d = 1$ and $\gamma_d = 0$, and $\omega_d = 1$ if $y_d = 0$ or $y_d = 1$ and $\gamma_d = 1$.

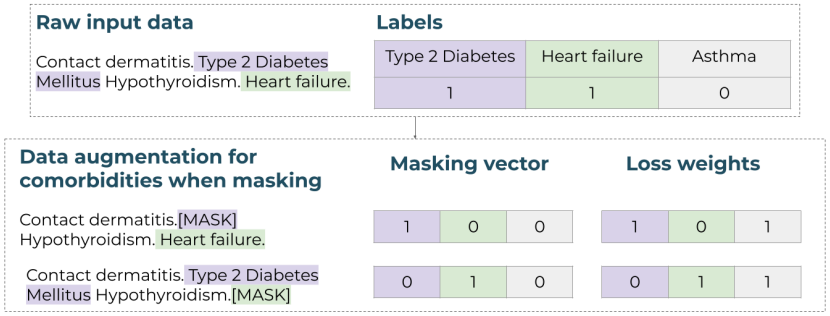


Figure 3: Example of an input sequence containing descriptions relevant to two positive phenotype labels in the target vector y. The input sequence is duplicated, and each phenotype is masked once and a loss weights vector is defined. Clinical terms relating to a directly associated phenotype d remain in the input data ($y_d = 1$ and $\gamma_d = 0$) but get assigned a loss weight $w_d = 0$ and so do not contribute to the loss function.

[1]Any patient data shown is simulated and does not represent data of real patients.

Positive weights. Because some diseases are very rare, whereas others are very common, our prediction classes are expected to be highly imbalanced in the practical setting. For cohort expansion, we want to increase recall while balancing a decline in precision. For $d \in \Delta$, let TN_d and TP_d denote the total number of negative and positive examples of d in the training set. We define the positive weight ρ_d as $\rho_d = TN_d/TP_d$, for all d $\in \Delta$.

Our loss function is defined as a mean-reduced binary cross-entropy loss function over phenotypes, where disease prevalence and comorbidities are handled with loss and positive weights:

$$\mathcal{L}^{(p)} = -\frac{1}{|\Delta|} \sum_{d \in \Delta} \omega_d^{(p)} (\rho_d y_d^{(p)} \log \sigma(z_d^{(p)}) + (1 - y_d^{(p)}) \log \sigma(1 - z_d^{(p)})), \tag{1}$$

where σ denotes the sigmoid function, ω_d the comorbidity-derived loss weight, ρ_d the positive weight, and $z_d^{(p)}$ the predicted probability for phenotype $d \in \Delta$ for sample (e.g. individual) p.

3 Experiments and Results

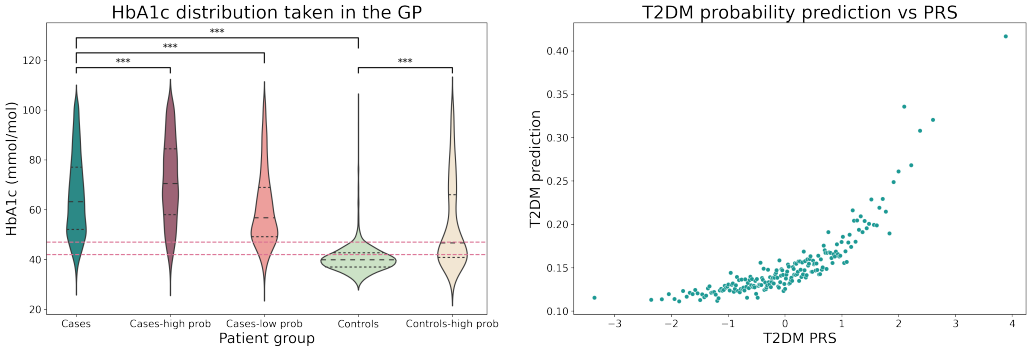
Data. The UK Biobank (UKBB) [Sudlow et al., 2015] is a large-scale research study of around 500k individuals. We restrict the data set to those that have entries in both hospital and GP records, reducing our cohort to 155k. To assess the quality of our model predictions, we choose four diseases that differ in terms of prevalence and clinical characteristics (Appendix C): Type 2 diabetes mellitus (T2DM), Heart failure (HF), Breast Cancer, and Prostate Cancer. We use validated phenotype definitions from CALIBER Kuan et al. [2019] to label patients with each of the diseases (Appendix C.1). We test the performance of our model on its ability to diagnose known cases, compare it to other methods, and evaluate associations of clinical features with predictions on T2DM.

sEHR-CE. We use the pre-trained language model PubMedBERT [Gu et al., 2022] as the encoder of the tokenised input sequences of clinical term descriptions. Since our input systematically differs from the general scientific text on which PubMedBERT was trained, we fine-tune on the masked-language modeling (MLM) task using the full UKBB cohort. The proposed model sEHR-CE uses the fine-tuned encoder and a fully connected linear layer as the decoder. To train on the multi-label classification task of outcome prediction, we split our data set into five equally sampled folds $f_0, ...f_4$ containing unique patients, and mask the data according to our strategy (Section 2.1, Figure C.1). We train a total of five models on three folds, holding back folds f_i for validation and $f_{(i+1)5}$ for testing for model i, $i = 1, \ldots, 5$ (Figure C.2). All results presented are predictions on the independent test set. For masking and training details, see Appendix C.2.

Baseline Models. We compare the performance of our model sEHR-CE to BEHRT [Li et al., 2020], which takes a tokenised sequence of clinical terms, age and position embeddings as input. Ontologies from hospital and GP records are mapped to CALIBER definitions, removing unmapped terms (more details in Appendix C.3). We train five such BEHRT models to predict an individual developing the four phenotypes on the same five data splits as sEHR-CE. Similarly, we train five sEHR-CE models restricted to CALIBER tokens (denoted sEHR-CE-codes) for comparison. Figure C.3 shows the distributions of predicted probabilites for all phenotypes across all methods. sEHR-CE shows the best performance across all four phenotypes in terms of recall at 0.5 and AUC on the test set (Table 1, Figure C.4). BEHRT performs slightly better than sEHR-CE-codes, indicating a benefit of adding visit position and age. Performance varies across phenotypes, presumably due to different clinical characteristics making some diseases easier to predict than others.

	T2D		HF		Breast Cancer		Prostate Cancer		Average	
	Recall	PRC	Recall	PRC	Recall	PRC	Recall	PRC	Recall, std	PRC
sEHR-CE-codes	0.64	0.70	0.76	0.60	0.42	0.45	0.36	0.26	0.55 ± 0.19	0.50
BEHRT	0.66	0.70	0.78	0.61	0.48	0.53	0.41	0.29	0.58 ± 0.17	0.53
sEHR-CE	**0.74**	0.74	**0.85**	0.69	**0.55**	0.55	**0.57**	0.47	$\mathbf{0.68 \pm 0.14}$	0.61

Table 1: Average and phenotype specific recall and AUCPR at 0.5 on the test sets.



(a) (b)

Figure 4: 4a HbA1c distribution across groups in the test sets with diabetes diganosis ranges shown as dashed lines. Significant p-values are indicated with *** (t-test, $\alpha = .001$). 4b Median T2DM prediction of individuals in the test sets grouped by the percentiles of the polygenic risk score.

3.1 Clinical evaluation on Type 2 Diabetes Mellitus

T2DM lends itself as a use case to qualitatively evaluate the predictions of missed cases, as it is a well-studied, slowly developing disease with varying disease severity. We used thresholds based on the 98th, 90th and 12th percentiles of sEHR-CE's predicted probabilities (Figure C.3) to define five different groups (Table C.1): regular cases, cases predicted with high probability, cases predicted with low probability, regular controls and controls predicted with high probability (missed cases).

Measures of Disease Severity. Haemoglobin A1c (HbA1c) is a blood biomarker used to diagnose and monitor diabetes. A level of 48mmol/mol or higher is considered diabetes; while a value between 42 and 48 mmol/mol is considered pre-diabetes. The input data did not include such biomarker data, so we use it here for evaluation. We aggregated HbA1c measurements taken in primary care (GP) of each individual by taking the 95-th percentile value. Figure 4a shows that cases predicted with high probability had the highest HbA1c mean levels. Cases identified with low probability were in the prediabetic range of HbA1c levels, indicating a less severe state. Missed cases had elevated HbA1c levels close to the prediabetic stage, representing individuals at risk of developing T2DM. We further investigated the association of the T2DM predicted probabilities with other measures of disease severity, expanded in Appendix C.4. Taken together, our results demonstrate that sEHR-CE's predicted probabilities of being diagnosed with T2DM are associated with disease severity.

Polygenic risk scores. Genetic risk for complex diseases like T2DM arises from many genetic changes that, when taken together, can increase an individual's risk of developing the disease, which can be defined by polygenic risk scores (PRS). Sinnott-Armstrong et al. [2021] developed PRS based on the UK Biobank participants for a set of diseases, including T2DM. We computed and standardised T2DM PRS across all individuals in our cohort [Lewis and Vassos, 2017]. Figure 4b demonstrates that higher predicted probability of T2DM was associated with a higher genetic risk.

4 Conclusion

We presented a data-driven method for cohort expansion based on language modelling. Our approach fuses primary and secondary care data via text, and we propose a data augmentation approach to allow for comorbidities in a patient's history. Our method predicts disease phenotype labels more accurately than non-text and single terminology approaches. We presented a high degree of evidence that our model identifies previously undiagnosed individuals that can extend the original cohort for downstream analysis. Future work will consider methods that are less restrictive on sequence length [Kitaev et al., 2020, Beltagy et al., 2020] and allow for irregular time steps [Shukla and Marlin, 2021] and age [Kazemi et al., 2019], as well as adding more data sources (e.g. medications).

References

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Sci. Rep.*, 10(1):7155, 2020.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021.

Abdelaali Hassaine, Dexter Canoy, Jose Roberto Ayala Solares, Yajie Zhu, Shishir Rao, Yikuan Li, Mariagrazia Zottoli, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Learning multimorbidity patterns from electronic health records using non-negative matrix factorisation. *Journal of Biomedical Informatics*, 112:103606, 2020.

Paul E Stang, Patrick B Ryan, Judith A Racoosin, J Marc Overhage, Abraham G Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann. Intern. Med.*, 153(9):600–606, 2010.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23, 2022.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. arxiv:1609.08144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388. Association for Computational Linguistics, 2019.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 2015.

Valerie Kuan, Spiros Denaxas, Arturo Gonzalez-Izquierdo, Kenan Direk, Osman Bhatti, Shanaz Husain, Shailen Sutaria, Melanie Hingorani, Dorothea Nitsch, Constantinos A Parisinos, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health*, 1(2):e63–e77, 2019.

Nasa Sinnott-Armstrong, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, Michael Wainberg, Hanna M Ollila, Tuomo Kiiskinen, Aki S Havulinna, James P Pirruccello, Junyang Qian, Anna Shcherbina, FinnGen, Fatima Rodriguez, Themistocles L Assimes, Vineeta Agarwala, Robert Tibshirani, Trevor Hastie, Samuli Ripatti, Jonathan K Pritchard, Mark J Daly, and Manuel A Rivas. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.*, 53(2):185–194, 2021.

Cathryn M Lewis and Evangelos Vassos. Prospects for using risk scores in polygenic medicine. *Genome Med.*, 9(1):96, 2017.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. 2020. arxiv:2004.05150.

Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time, 2019. arXiv:1907.05321.

Jette Henderson, Joyce C. Ho, Abel N. Kho, Joshua C. Denny, Bradley A. Malin, Jimeng Sun, and Joydeep Ghosh. Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 214–223, 2017.

Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6(1): 26094, 2016.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *PMLR*, pages 301–318, 2016a.

Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. DeepCare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining*, pages 30–41. Springer International Publishing, 2016.

Ahmed M. Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1495–1504. Association for Computing Machinery, 2016b.

Spiros Denaxas, Pontus Stenetorp, Sebastian Riedel, Maria Pikoula, Richard Dobson, and Harry Hemingway. Application of clinical concept embeddings for heart failure prediction in UK EHR data. *NIPS ML4H: Machine Learning for Health*, 2018.

Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. Medical concept embedding with time-aware attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3984–3990. AAAI Press, 2018.

Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. TAPER: Time-aware patient EHR representation. *IEEE journal of biomedical and health informatics*, 24(11):3268—3275, 2020.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *PMLR*, pages 183–203, 2022.

Martin Chapman, Shahzad Mumtaz, Luke V Rasmussen, Andreas Karwath, Georgios V Gkoutos, Chuang Gao, Dan Thayer, Jennifer A Pacheco, Helen Parkinson, Rachel L Richesson, Emily Jefferson, Spiros Denaxas, and Vasa Curcin. Desiderata for the development of next-generation electronic health record phenotype libraries. *Gigascience*, 10, 2021.

WHO. Who. diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed: 19-08-2022.

Medscape. Type 2 diabetes mellitus treatment & management. https://emedicine.medscape.com/article/117853-treatment, Accessed: 03-08-2022.

João Pedro Ferreira, Sarah Kraus, Sharon Mitchell, Pablo Perel, Daniel Piñeiro, Ovidiu Chioncel, Roberto Colque, Rudolf A de Boer, Juan Esteban Gomez-Mesa, Hugo Grancelli, Carolyn S P Lam, Antoni Martinez-Rubio, John J V McMurray, Alexandre Mebazaa, Gurusher Panjrath, Ileana L Piña, Mahmoud Sani, David Sim, Mary Walsh, Clyde Yancy, Faiez Zannad, and Karen Sliwa. World heart federation roadmap for heart failure. *Glob. Heart*, 14(3):197–214, 2019.

Diana Ly, David Forman, Jacques Ferlay, Louise Brinton, and Michael Cook. An international comparison of male and female breast cancer incidence rates. *International journal of cancer*, 132, 2013.

Prashanth Rawla. Epidemiology of prostate cancer. *World Journal of Oncology*, 10:63–89, 2019.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. On the stratification of multi-label data. In *ECML/PKDD*, 2011.

A Related Work

ML approaches have been applied to EHR data either using a cross-sectional matrix of diagnosis terms (e.g. an ICD-10 term) [Henderson et al., 2017, Miotto et al., 2016] or sequential data as input [Choi et al., 2016a, Pham et al., 2016, Alaa and van der Schaar, 2019, Choi et al., 2016b, Denaxas et al., 2018, Cai et al., 2018, Darabi et al., 2020]. In the former, methods are blind to the order in which diagnoses occur and subsequently how a patient's disease profile develops. In the latter, most methods do not learn embeddings for the full sequence of diagnoses in a patient's medical history, and instead learn embeddings per diagnosis term or, at most, a short sequence of terms whether they are utilising LSTMs [Choi et al., 2016a], RNNs [Pham et al., 2016], CNNs [Nguyen et al., 2017], or transformer models [Li et al., 2020]. To be able to deal with heterogeneous ontologies from different EHR data sources, all of these models rely on noisy and often lossy mappings across ontologies or on phenotyping algorithms, eg. manually curated groupings of ontology terms, such as CALIBER [Kuan et al., 2019].

In contrast, our method uses the textual description of terms to learn representations across the full sequence of diagnoses in a patient's history. Recent work by Hur et al. [2022] aims to unify EHR records by learning description-based embeddings from multiple data sources. Our work was developed in parallel independently and addresses the specific use case of cohort expansion, instead of merely providing examples of potential downstream applications to assess improvements of predictive power.

B Method details and figures

B.1 Fusing ontologies via text

We consider all EHR data sources with their ontologies where each concept has a textual descriptor. For example, the ontologies of GP and hospital records are made up of clinical terms (e.g. Read version 2/ Clinical Terms Version 3 and ICD9/ICD10 codes, respectively) and their description. For each EHR source and associated ontology $a \in \mathcal{A}$, we denote the set of concepts (e.g. clinical terms in the case of GP or hospital records) within this ontology as Θ_a and the set of text descriptions as Ξ_a. The total vocabulary of concepts and text descriptions across all ontologies is denoted by $\Theta = \bigcup_{a \in \mathcal{A}} \Theta_a$ and $\Xi = \bigcup_{a \in \mathcal{A}} \Xi_a$, respectively. For each patient, we define their full clinical history through time and across sources as the sequence of time-indexed concepts as $(\theta_1, \ldots, \theta_t)$, $\theta_i \in \Theta$, $i = 1, \ldots, t$.

The premise of our work relies on the assumption that for every concept $\theta \in \Theta$, there exists a unique text description $\xi_\theta \in \Xi$. For example, under the ICD-10 terminology, the alphanumeric code E11.9 has the associated description 'Type 2 diabetes mellitus without complications'. Thus the clinical history of each patient can be uniquely represented by the concatenation of sequences of clinical descriptions $(\xi_{\theta_1}, \ldots, \xi_{\theta_t})$, $\xi_{\theta_i} \in \Xi$, $i = 1, \ldots, t$, ordered in time. Figure B.1 shows an example of fusing primary care and hospitalization records.

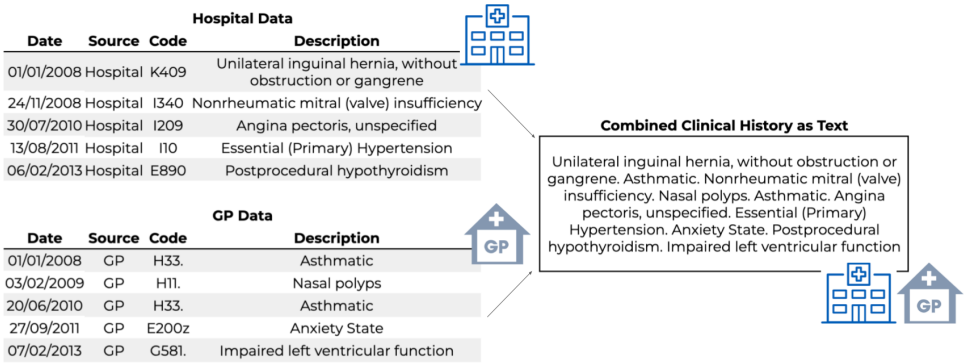


Figure B.1: Fusing primary (GP) and secondary (hospital) data into a single paragraph.

B.2 Label Generation: Oracle Feature Tagging for Disease Phenotyping

Given a set of clinical terms Θ and text descriptions Ξ, we rely on external oracles to assign labels to a given set of target phenotypes Δ. We assume that for each phenotype $d \in \Delta$ there exists a mapping $\mathbf{1}_d : \Theta \times \Xi \to \{0, 1\}$, $(\theta, \xi_\theta) \mapsto \delta_d$ indicating whether the presence of d can be inferred from the clinical term and its description. An aggregated phenotype label of 1 is assigned, if $\mathbf{1}_d(\theta, \xi_\theta) = 1$ for any of the term-description pairs (θ, ξ_θ) in the input sequence, and 0 otherwise (Figure 1). Here, Δ is a set of disease phenotypes, that can be taken from disease-specific phenotyping algorithms [Chapman et al., 2021], such as CALIBER [Kuan et al., 2019], which consists of manually-curated sets of clinical terms across primary and secondary care ontologies for defining 308 chronic and acute disease phenotypes.

B.3 Model Design

Figure B.2 shows a diagram of sEHR-CE.

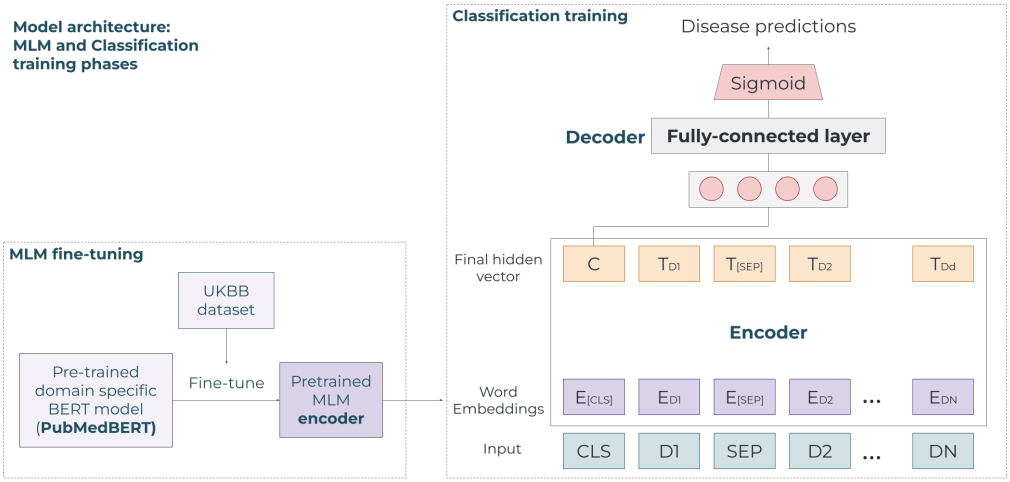


Figure B.2: Diagram of the sEHR-CE model. The model is fine-tuned on the MLM task. We then use the pre-trained encoder and train sEHR-CE. The input tokens are first encoded and the hidden vector of the [CLS] token is passed to the decoder, a fully connected linear layer. The output is passed through a sigmoid function to generate probabilities for each phenotype.

C Experiment design and model predictions

To assess the quality of our model predictions, we chose four diseases that differ in terms of prevalence and clinical characteristics. Type 2 diabetes mellitus (T2DM) is one of the most prevalent chronic diseases worldwide [WHO], and patients are primarily diagnosed and managed in primary care (GP) [Medscape]. Heart failure (HF) is one of the main causes of death in the older population [Ferreira et al., 2019], and its acute manifestations are treated in hospital care. Malignant neoplasms of the breast and of the prostate are both less prevalent diseases almost exclusively present in only biologically females or males, respectively [Ly et al., 2013, Rawla, 2019]. We test the performance of our model on its ability to diagnose known cases, compare it to other methods, and evaluate associations of clinical features with the predictions on T2DM with available orthogonal data.

C.1 Data Processing

The UK Biobank (UKBB) [Sudlow et al., 2015] is a large-scale research study of around 500k individuals between the ages of 40 and 54 at the time of recruitment. It includes rich genotyping and phenotyping data, both taken at recruitment and during primary and secondary care encounters (GP and hospital). We use patient records from GP and hospital visits in the form of code ontologies

Read version2/ Clinical Terms Version 3, and ICD-9/10 together with their textual descriptions. We restrict the data set to individuals that have entries in both hospital and GP records, which reduces our cohort to $154,668$ individuals. Requiring individuals to have entries in their GP records reduces bias towards acute events that usually present in hospitals, but we note that removing individuals without any hospital records may still bias the data towards more severe cases. We use CALIBER, previously validated phenotype definitions from Kuan et al. [2019] to label patients with T2DM, HF, malignant neoplasm of the breast, and malignant neoplasm of the prostate.

A patient can be admitted to the hospital for multiple days. We treat an entire hospital admission as one point in time using the admission date, and only keep unique ICD-10/ICD-9 codes for each visit. We aggregate visits that are less than a week apart into one visit keeping only unique codes. This approach removes repeated codes, thus avoiding redundancy and reducing sequence length.

Only patients with at least 5 clinical terms present in their clinical history are included to allow for sufficient information for any predictions, reducing the cohort to a final $129,932$ individuals. We use phenotype definitions from CALIBER [Kuan et al., 2019] to label patients with T2DM, HF, malignant neoplasm of the breast, and malignant neoplasm of the prostate. Each phenotype definition consists of a list of ICD10 and Read2, Read3 codes and their children.

The usage of PubMedBERT restricts the length of input sequences we can use. To avoid excluding relevant clinical information by truncating the input sequences, we break up patient histories longer than the limit into multiple input sequences of smaller length with the same target vector.

C.2 Model Training

We use the pre-trained language model PubMedBERT [Gu et al., 2022] as the encoder of the tokenised input sequences of clinical term descriptions. Since our input systematically differs from the general scientific text on which PubMedBERT was trained, we fine-tuned on the masked-language modeling (MLM) task (Figure B.2, Figure C.1), by masking words (e.g. descriptions) at random following the original BERT paper [Devlin et al., 2019]. The model, fine-tuned using the full UKBB cohort of $138,079$ patients, was trained with early stopping for 5 epochs with a batch size of 32 and a learning rate of 4×10^{-5} using gradient descent with an AdamW optimizer, and weight decay of 0.01. The output dimension of the encoder was 768.

The proposed model sEHR-CE is using the fine-tuned encoder and a fully connected linear layer as the decoder. To train on the multi-label classification task of outcome prediction, we split our data set into five equally sampled folds $f_0, ...f_4$ containing unique patients, using a stratified sampling method to maintain the same phenotype proportion in every split [Sechidis et al., 2011], and mask the data according to our strategy (Section 2.1, Figure C.1).

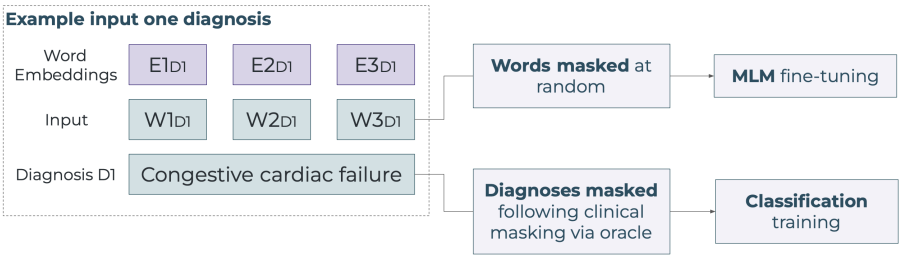


Figure C.1: Example of input diagnosis and different masking strategies for MLM fine-tuning and classification training. A description is encoded with multiple word embeddings. For MLM fine-tuning, words are masked at random; for classification training, whole descriptions are masked using clinical masking strategy described in Section 2.1.

We train a total of five models for 3 epochs on three folds, holding back folds f_i for validation and $f_{(i+1)5}$ for testing for model i, $i = 1, \ldots, 5$ (Figure C.2). This maintains a $60/20/20$ training, validation and testing split overall while providing us with enough training and testing examples. All results presented are predictions of each model on its respective independent test set. We used a learning rate of 10^{-5}, and a warm-up proportion of 0.25. Performance was monitored every 0.25 epochs on the validation fold.

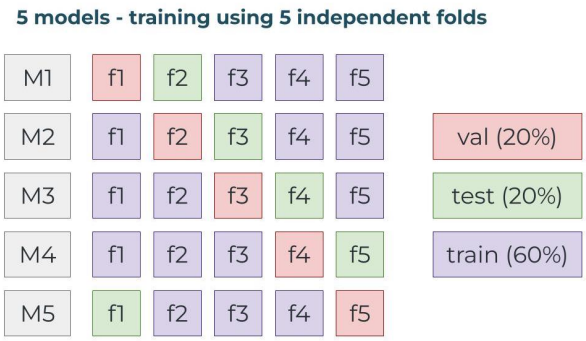


Figure C.2: Five models trained using five independent folds from the total data set using stratified sampling.

C.3 Model evaluation

As explained in 3, we compare the performance of our model sEHR-CE to BEHRT [Li et al., 2020]. BEHRT takes a tokenised sequence of clinical terms, age and position embeddings as input. Ontologies from hospital and GP records are mapped to CALIBER definitions [Kuan et al., 2019], removing unmapped terms. Phenotype definitions in CALIBER include different categories (for example, 'diabetes' contains categories 'type 1' and 'type 2'), that were ignored by the original BEHRT publication, so we expanded the token set to define a token per CALIBER phenotype and category. A transformer model is pre-trained to predict masked tokens before it is trained to predict a set of possible diagnoses an individual may develop given the input sequence. Similarly, we trained five sEHR-CE models restricted to CALIBER tokens (denoted sEHR-CE-codes) for comparison. All results presented are predictions of each model on its respective independent test set.

Figure C.3 shows the predicted probabilities for cases and controls across all phenotypes and models, and figure C.4 shows the AUPRC curves for each phenotype and method in the test sets. To avoid inclusion of too many false positives, we defined missing cases as those controls with predicted probability in the 98th percentile.

C.4 Expanded Qualitative Evaluation of T2DM Diagnosis Prediction

Five groups were used to evaluate the model's T2DM diagnosis predictions. Table C.1 shows the percentiles of sEHR-CE's predicted probabilities to define each group, along with each size.

Patient group	Size
Cases	16431
Controls	113501
Controls with high probability (p>=0.85, 98th percentile)	2020
Cases with high probability (p>=0.985, 90th percentile)	2072
Cases with low probability (p<=0.25, 12th percentile)	2343

Table C.1: Case and control cohorts and groups of interest based on sEHR-CE's predicted probability for T2DM in the test sets.

We then investigated the association of predicted probability and several proxies of disease severity: number of GP and hospital admissions, survival and risk of cardiovascular disease.

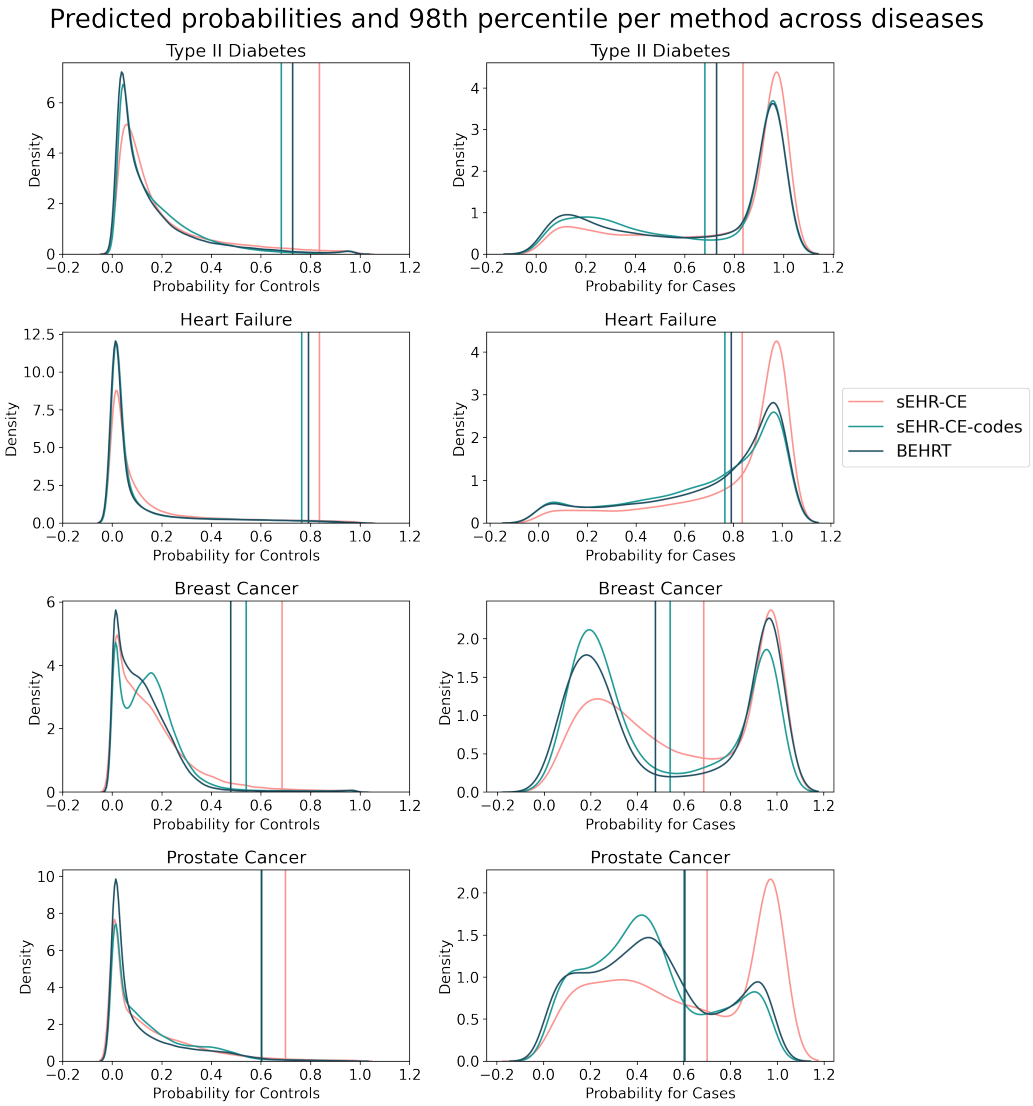


Figure C.3: Predicted probabilities for cases and controls in test sets across all phenotypes. Vertical lines indicate the 98th percentile threshold.

C.4.1 Number of GP visits and hospital admissions

As expected, both cases and controls with a high predicted probability of a T2DM diagnosis, exhibit a slightly higher number of GP and hospital visits than the other groups (Figure C.5), indicating that they are experiencing a more severe form of T2DM requiring care. This is particularly higher in the case of hospital visits, indicating patients experiencing acute events: both cases and controls with a high predicted probability visit a hospital approximately 10 times more often than their low probability counterparts.

Although the model was not given information from which data source the input data was coming from, this analysis indicates that it has learned to associate acute events with disease severity.

C.4.2 Survival analysis

To compare survival across different groups of individuals, we use the Kaplan-Meier estimator with all-cause mortality as the endpoint with right-censored data accounting for individuals without any event occurrence since the last follow-up. Both cases and controls with high predicted probabilities

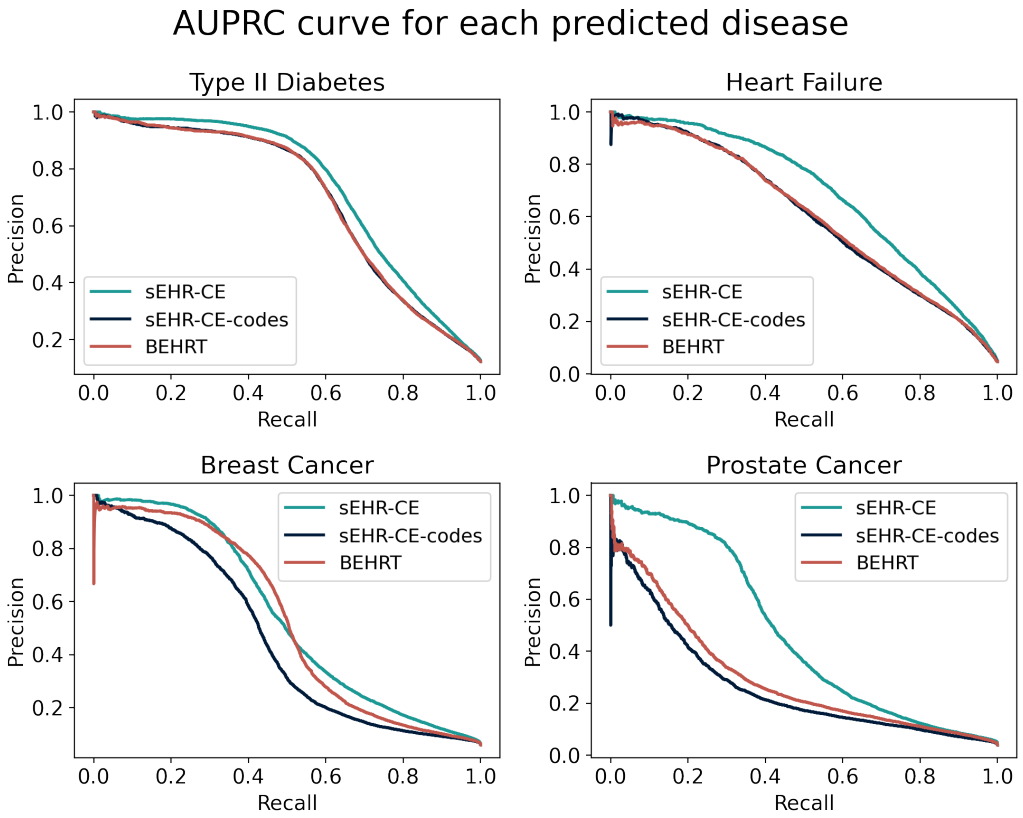


Figure C.4: AUPRC curves for each phenotype in the test sets.

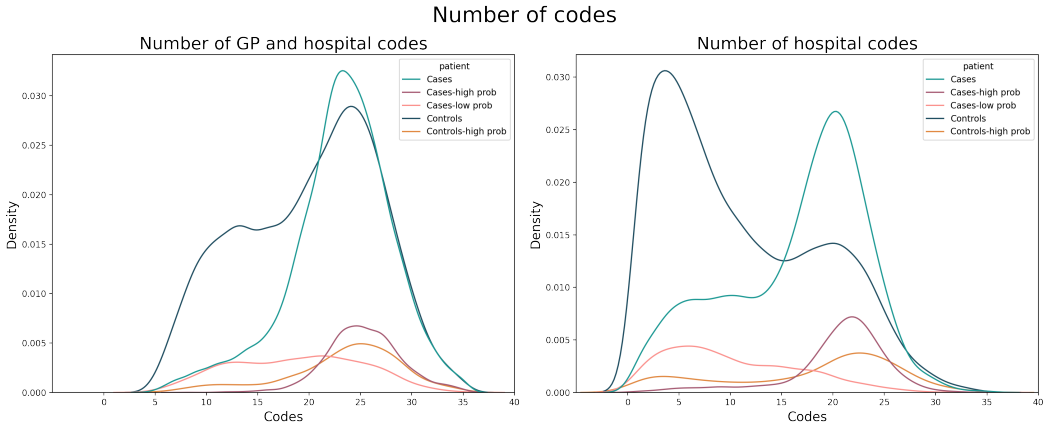
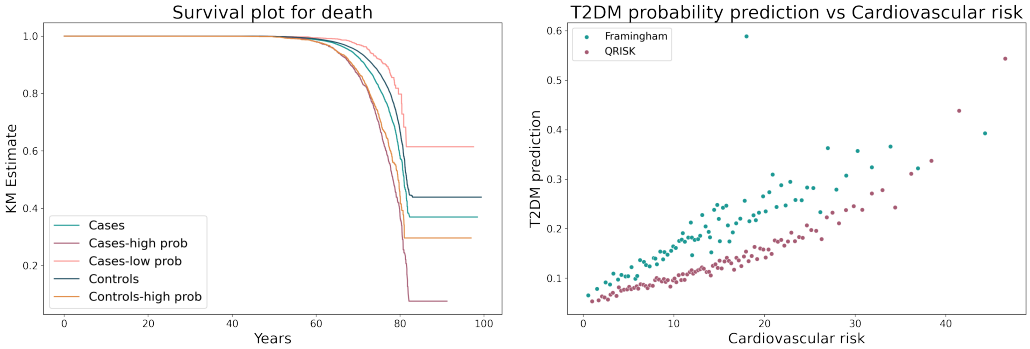


Figure C.5: Distribution of number of hospital and GP codes per patient group. Matched controls on sex.

had the lowest survival, followed by general cases, controls and finally cases with low predicted probability (Figure C.6a), indicating that the model's predicted probability is associated with survival.

C.4.3 Cardiovascular Risk

T2DM is a known risk factor and comorbidity of cardiovascular disease, which, in turn, is the most prevalent cause of death in T2DM patients. The GP records contain Framingham and QRISK3 scores; these are two scores that assess an individual's risk of developing cardiovascular disease



(a) (b)

Figure C.6: C.6a Death survival plots for different patient groups. C.6b Framingham and QRISK cardiovascular scores vs T2DM probability prediction. Each point represents the median of each percentile of cardiovascular risk.

within the next 10 years, based on several coronary risk factors. The Framingham score is derived from an individual's age, gender, total cholesterol, high-density lipoprotein cholesterol, smoking habits, and systolic blood pressure, whereas the QRISK score, which is almost exclusively used now, extends this score with additional factors such as body mass index, ethnicity, measures of deprivation, chronic kidney disease, rheumatoid arthritis, atrial fibrillation, diabetes mellitus, and anti-hypertensive treatment. Both cases and controls with high predicted probability of having T2DM had a higher risk of developing cardiovascular disease compared to their low predicted probability counterparts (Figure C.6b) indicating that the model has learned to associate the risk of developing both diseases at the same time.