Deep representation learning for clustering longitudinal survival data from electronic health records

Jiajun Qiu[1], Yao Hu[2], Frank Li[3], Abdullah Mesut Erzurumluoglu[2], Ingrid Braenne[2], Charles Whitehurst[4], Jochen Schmitz[5], Johann de Jong[1*]

[1]Statistical Modeling, Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88400 Biberach an der Riβ, Germany

[2]Human Genetics, Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88400 Biberach an der Riβ, Germany

[3]CB I&R, Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88400 Biberach an der Riβ, Germany

[4]Drug Concept Discovery I&R, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88400 Biberach an der Riβ, Germany

[5]Disease Positioning I&R, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88400 Biberach an der Riβ, Germany

*Corresponding author

Abstract:

Precision medicine can be defined as providing the right treatment to the right patient at the right time, and it requires the ability to identify clinically relevant patient subgroups with high accuracy. The increasing availability of large-scale electronic health records (EHR) datasets has provided major opportunities for artificial intelligence and machine learning in mining such complex datasets for identifying novel disease subtypes. However, disease subtypes often exist in the context of certain disease-relevant risk events, and current efforts have been limited by analyzing clustering and event risk independently, resulting in subgroups that still display great heterogeneity in event risk and/or underlying molecular mechanisms.

To address this problem, we developed TransVarSur (Transformer Variational Survival modeling). TransVarSur integrates a Transformer-based Gaussian mixture variational autoencoder with time-to-event modeling to capture complex relationships between cluster-specific EHR trajectories and survival times. We validated TransVarSur by showing superior performance relative to baseline methods, on both synthetic and real-world benchmark datasets with known ground-truth clustering. We then applied TransVarSur to 1908 Crohn's disease patients from the UK Biobank and successfully identified four clusters displaying both divergent EHR trajectories and divergent progression towards the risk event intestinal obstruction. A further analysis of the clusters revealed known clinical and genetic factors relevant in Crohn's disease and progression to intestinal obstruction.

In conclusion, we demonstrated TransVarSur’s ability to accurately stratify a patient population into clinically and genetically relevant, risk-associated subgroups. Hence, it can be a powerful tool in the development of precision medicine approaches.

INTRODUCTION

There has been a notable shift in the healthcare sector towards digitizing patient information, with electronic health records (EHRs) emerging as the new norm. As of 2018, the adoption rate of EHR systems has surpassed 84% and 94% in the US and UK, respectively [1,2]. EHR systems offer a comprehensive and easily accessible source of patient data, typically gathering data from millions of individuals over many years, encompassing various sources (such as primary and secondary care) and modalities (such as diagnoses, medications, and lab tests). The extensive nature of EHRs makes them a valuable resource for healthcare research, enabling more accurate modeling of patients and their disease risk, onset, and progression. However, the sheer size and complexity of EHR data inevitably pose challenges to modeling efforts, necessitating the development of sophisticated algorithms and data processing methods. Rapid advancements in the field of artificial intelligence (AI) and machine learning (ML) have had a profound impact on a wide range of industries and provide many opportunities for improving healthcare as well [3,4]. Specifically, machine learning has shown great promise in EHR research due to its ability to uncover hidden patterns and trends in such complex datasets. AI/ML methods have been successfully used for addressing a range of research questions, including modeling disease progression [4], predicting patient outcomes [5], and identifying novel disease subtypes [6].

Here, we explore the application of AI/ML methods to the problem of EHR-based patient clustering in the context of disease-associated risk events. Patient clustering is an important concept in the field of precision medicine. Precision medicine aims to provide the right treatment, to the right patient at the right time, by utilizing individual patient characteristics to guide clinical decision-making, instead of population-wide averages of patient characteristics [7]. Patient clustering supports the development of precision medicine approaches by detecting patterns and trends within a certain patient population of interest, which can serve as the basis for the identification of novel disease subtypes. By studying the causal molecular mechanisms of these disease subtypes, more targeted and personalized therapeutic approaches can be developed. Disease subtyping is often relevant in the context of certain risk events [8]. For example, why do some patients with Crohn's disease, a subtype of inflammatory bowel disease (IBD) progress to intestinal stricture (a narrowing of the intestines due to the formation of scar tissue and muscular hypertrophy), and others do not? As Crohn’s disease is a multifactorial disease, it is likely that there are multiple mechanisms associated with progression towards intestinal obstruction [9]. Typically, one of the following two approaches is applied for elucidating how patient subgroups correlate with event risk: (1) Start with identifying patient clusters, and then analyze the risk event within each of the clusters [10]. This approach has inherent limitations because resulting clusters are not guaranteed to correlate to the event risk. (2) Start with stratifying patients by the risk event, and then identify subgroups within the risk strata [11]. The main limitation here is that it can be difficult to identify patient subgroups with differentiating generative mechanisms. In other words, a high-risk patient subgroup could exhibit significant heterogeneity in causal molecular mechanisms, and patient subgroups with comparable survival outcomes could have varying responses to identical treatments. [12]. Hence, to enhance the identification of novel disease subtypes in the context of risk events of interest, patient clustering should be integrated with risk modeling (time-to-event analysis).

Several studies have previously explored the combination of clustering and risk modeling. In one of the earliest efforts, Bair & Tibshirani *et al*. introduced SSC (Semi-Supervised Clustering), which initially selects features based on univariate Cox regression hazard scores and then conducts k-means clustering using the selected features [13]. More recently, Chapfuwa *et al.* presented SCA (Survival Cluster Analysis), which uses a neural network to map

covariates to a latent space that is encouraged to follow a mixture of truncated Dirichlet processes [14]. Additionally, Nagpal *et al.* developed DSM (Deep Survival Machines), a deep neural network for learning patient representations while regularizing these towards a mixture of Weibull distributions [15]. The authors later extended this framework to RDSM (Recurrent Deep Survival Machines), for modeling longitudinal data using recurrent neural networks [16]. However, none of these approaches directly integrated risk modeling with clustering. In all cases, the resulting clusters are purely driven by survival outcome (Table 1) and will suffer from the limitations we previously outlined. These limitations were eventually addressed by the introduction of VaDeSC (Variational deep survival clustering) [17], which directly integrates clustering and risk modeling by expressing cluster-specific associations between covariates and survival times in a single semi-supervised model, using a Gaussian mixture variational framework. Although VaDeSC provided a major advance in clustering survival data, it does not consider the longitudinal aspects of patient modeling. Insights into patients' disease history and progression are however of great importance for developing a more comprehensive disease understanding and eventually developing more effective and personalized therapies [18,19].

In this work, we build upon recent advances in using language modeling technology for analyzing EHR sequences [4] and present a novel method for clustering longitudinal survival data: TransVarSur (**Trans**former **Var**iational **Sur**vival modeling). TransVarSur extends the VaDeSC approach by modeling longitudinal patient data using an autoencoding transformer architecture. In this study, we demonstrate TransVarSur's ability to capture statistical interactions between cluster-specific disease trajectories and survival times, enabling it to discover novel clinically relevant patient subgroups.

Table 1: Conceptual comparison of TransVarSur and related approaches. Here, t denotes survival time, x denotes the input features, z corresponds to the latent representations. PH: Proportional Hazard. SSC: Semi-Supervised Clustering. SCA: Survival Cluster Analysis. DSM: Deep Survival Machines. RDSM: Recurrent Deep Survival Machines. VaDeSC: Variational deep survival clustering.

	K-means + Cox PH	**SSC**	**SCA**	**DSM**	**RDSM**	**VaDeSC**	**TransVarSur**
Predicts t?	No	No	Yes	Yes	Yes	Yes	Yes
Learns z?	No	No	Yes	Yes	Yes	Yes	Yes
Models interactions between x and t?	No	No	No	No	No	Yes	Yes
Longitudinal model?	No	No	No	No	Yes	No	Yes

RESULTS

TransVarSur architecture

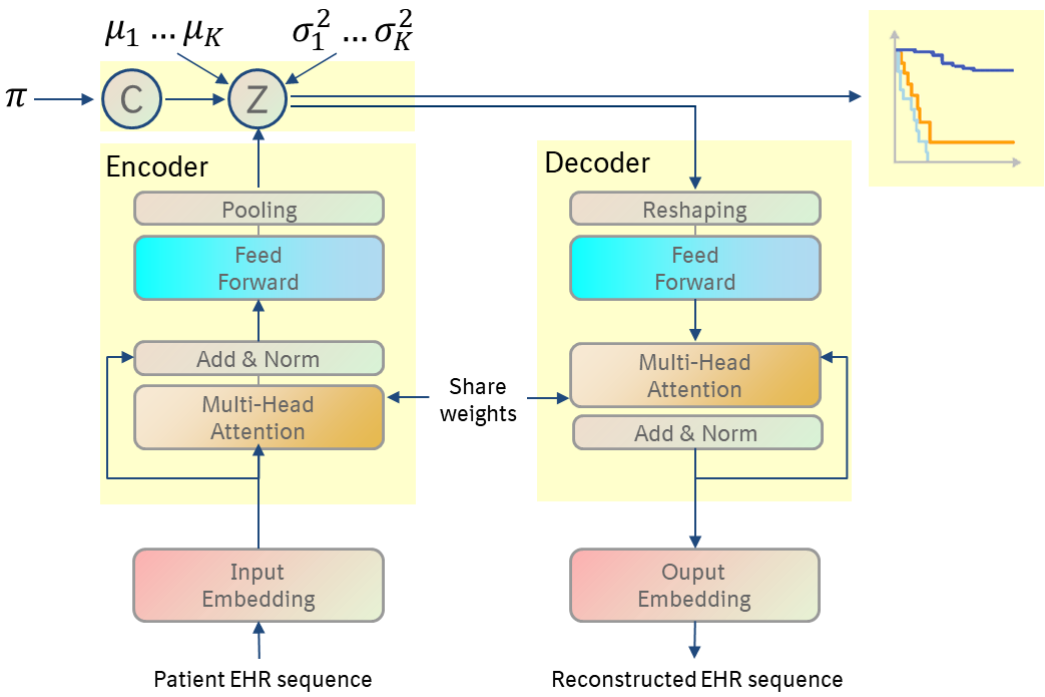

Figure 1. Architecture of TransVarSur. First, an embedding is computed for a patient's EHR sequence. This embedding serves as input for multiple transformer blocks, where the final pooling layer generates the latent representation Z for the patient. Z is regularized towards a Gaussian mixture distribution by including a variational term in the loss function. Z is then used to predict the time-to-event, as well as passed to the transformer decoder for reconstructing the patient's EHR sequence. Weights are shared between the encoder and decoder. For more details, please refer to the Methods section.

Position embedding	0	0	0	1	2	3	3	4
Type: Primary/Hospital	Primary	Primary	Primary	Hospital	Hospital	Hospital	Hospital	Hospital
Age	36.33	36.33	36.33	42.25	48.00	53.58	53.58	60.22
ICD10: Block	R50-R69	K50-K52	O30-O48	E65-E68	Z80-Z99	I10-I15	I32-I52	E50-E64
ICD10: category	R63	K50	O43	E66	Z90	I10	I51	E53
ICD10: subcategory	R635	K509	O431	E660	Z904	I110	I516	E538
	Diag. 1	Diag.2	Diag.3	Diag.4	Diag.5	Diag.6	Diag.7	Diag.8

Figure 2. The embedding structure. Simulated example of an EHR sequence for a single individual with 8 diagnoses spread across 5 primary or hospital care visits (0 - 4) embedded at three levels of the ICD10 ontology. Diag.: Diagnosis.

Our proposed method, TransVarSur, takes a patient's disease history (diagnosis sequence) as an input and maps it into a latent representation z using a transformer-based variational autoencoder (VAE) with a Gaussian mixture prior (Figure 1). The survival outcome is modeled using a mixture of Weibull distributions with cluster-specific parameters β. The parameters of the Gaussian mixture and Weibull distributions are jointly optimized using the EHR sequences and survival outcomes. Note henceforth we use the terms survival modeling, risk modeling and time-to-event modeling interchangeably.

For the input embedding layer, each diagnosis is represented by a combination of six distinct embeddings (Figure 2): three for the ICD10 code (subcategory, category, and block), one for age, one for type, and one for position. The three-level embedding for ICD10 diagnoses is designed to capture the hierarchical structure inherent in the ICD10 coding system. The age embedding represents the patient's age at the time of diagnosis and can additionally assist the model in understanding the temporal gaps between diagnoses. The type embedding differentiates between diagnoses derived from primary care data and those from hospital data. The position embedding, representing visits, establishes the relative placement of diagnoses within the EHR sequence, allowing the network to recognize positional relationships among diagnoses. Diagnoses originating from the same visit will have identical position embeddings.

Details around the model architecture and loss function can be found in the Methods section.

TransVarSur outperforms baseline methods on simulated benchmark

As a first step in validating TransVarSur, we assessed performance on a benchmark dataset simulated using a TransVarSur decoder (see Method section). In addition to evaluating the cluster predictions using the balanced accuracy (ACC), normalized mutual information (NMI) and adjusted Rand index (ARI), we evaluated the time-to-event predictions using concordance index (CI). Note that due to the noise in the data generating process, achieving perfect performance was impossible.

Table 2. Performance on simulated benchmark data. Comparison between TransVarSur and the other methods used for clustering survival data, in terms of balanced accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and concordance index (CI). Reported ± is one standard error, and significance of the difference (p-val < 0.05) between TransVarSur and the other methods is indicated by an asterisk.

Method	ACC	NMI	ARI	CI
***k*-means+Cox PH**	0.334±0.002*	0.0003±0.0001*	-0.0003±0.0003*	0.505±0.001*
SSC	0.3334±0.0004*	0.0003±0.0002*	0.00005±0.00050*	
SCA	0.336±0.007*	0.03±0.02*	0.006±0.040*	0.639±0.006*
DSM	0.3327±0.0004*	0.0005±0.0005*	-0.0005±0.001*	0.49±0.01*
RDSM	0.502±0.002*	0*	0*	0.500±0.001*
VaDeSC	0.57±0.05*	0.37±0.04*	0.44±0.03*	0.79±0.02*
TransVarSur	0.64±0.04	0.72±0.04	0.66±0.07	0.77±0.01

As can be seen, TransVarSur significantly outperformed all other methods in retrieving the ground truth clusters from the benchmark data (Table 2). The next best performing method

was VaDeSC, the method most closely related to TransVarSur. The main difference between the two methods is that TransVarSur incorporates a transformer-based architecture for modeling the longitudinal nature of the EHR data, whereas VaDeSC relies on a simple fully connected neural network with TF-IDF features. As expected, the transformer architecture helped TransVarSur to achieve much better performance on the longitudinal clustering task than VaDeSC (TransVarSur: ACC = 0.78±0.04, NMI = 0.72±0.04 and ARI = 0.66±0.07; VaDeSC: ACC = 0.67±0.02, NMI = 0.37±0.04 and ARI = 0.44±0.03). Importantly, TransVarSur achieved its superior clustering performance while hardly sacrificing performance on the risk prediction task. VaDeSC only showed marginally better performance on survival prediction than TransVarSur (CI = 0.79±0.02 for VaDeSC vs. CI = 0.77±0.01 for TransVarSur, p-val = 0.007). This marginal performance difference can potentially be explained by the fact that TransVarSur was trained for 200 epochs compared 160 for VaDeSC because its transformer architecture has many more parameters than VaDeSC's simple feedforward neural network. During these 40 additional epochs, TransVarSur may have mildly overfit on the survival times relative to the diagnosis data, as compared to VaDeSC. It is however important to note that, should such differences exist and be relevant in a given setting, they are easily mitigated by weighting the different components of the loss function (e.g. [20]).

The remaining models (k-means+Cox PH, SSC, SCA, DSM and RDSM) performed no better than random at retrieving the ground truth clusters. Whereas RDSM clearly stood out as the third-best performing on the risk prediction task, it did perform worse than VaDeSC. This is interesting, because RDSM explicitly models the EHR data as sequences, whereas VaDeSC does not. However, as explained in the introduction, RDMS's clusters are solely driven by survival times. This can explain the observed difference between VaDeSC and RDSM on the risk prediction task, and again highlights the importance of modeling the interactions between the diagnosis sequences and the survival times.

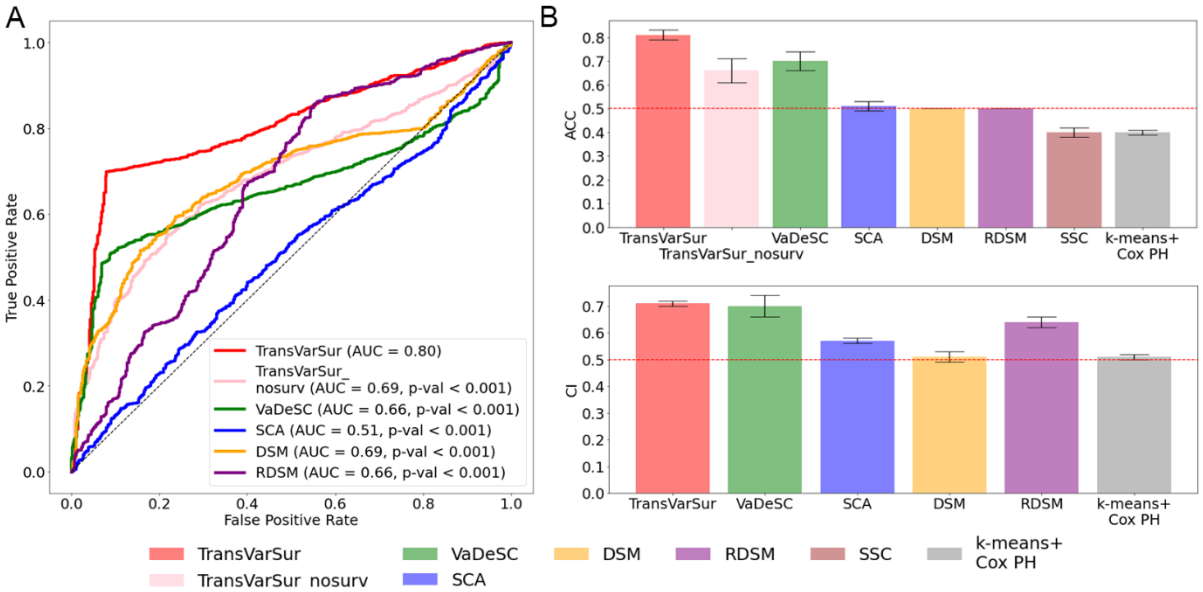

Figure 3. Performance on simulated benchmark data. Comparison between TransVarSur and the other methods used for clustering survival data. TransVarSur_nosurv represents TransVarSur trained without risk loss. A) Performance on retrieving the ground-truth clustering, in terms of the area under the receiver-operating characteristic (ROC), with p-values for the significance of the difference between TransVarSur and the other methods; B) Performance on retrieving the ground-truth clustering in terms of balanced accuracy (ACC), with 0.5 for random performance.; C) Performance on time-to-event prediction, in terms of concordance index (CI), with 0.5 for random performance.

Table 3. Performance on T1D/T2D benchmark data. Comparison between TransVarSur and the other methods used for clustering survival data, in terms of balanced accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and concordance index (CI). Reported ± is one standard error, and significance of the difference (p-val < 0.05) between TransVarSur and the other methods is indicated by an asterisk.

Method	ACC	NMI	ARI	CI
***k*-means+Cox PH**	0.40±0.01*	0.06±0.01*	-0.09±0.01*	0.51±0.01*
SSC	0.40±0.02*	0.06±0.02*	-0.087±0.007*	
SCA	0.51±0.02*	0.003±0.003*	0.01±0.01*	0.57±0.01*
DSM	0.50±6.62e-05*	7.2e-8±1.44e-7*	-0.00015±0.0003*	0.37±0.02*
RDSM	0.50*	0*	0*	0.65±0.02
VaDeSC	0.70±0.04*	0.12±0.07*	0.11±0.09	0.70±0.04
TransVarSur_nosurv	0.66±0.05*	0.06±0.03*	0.07±0.06*	
TransVarSur	0.81±0.02	0.24±0.04	0.22±0.04	0.71±0.01

TransVarSur outperforms baseline methods on a T1D/T2D benchmark

We next assessed the performance of TransVarSur and the other methods on a specifically designed real-world use case from UK Biobank data with known ground-truth cluster labels: distinguishing 494 type 1 diabetes (T1D) patients from 1830 type 2 diabetes (T2D) patients in their progression towards retinal disorders. Note that the T1D and T2D labels were not used for clustering, and thus also the age at first T1D or T2D diagnosis was not available to the model. The main observations strongly mirror those made from the simulation benchmark. TransVarSur outperformed the other methods at retrieving the ground truth clustering (Figure 3 and Table 3; AUC = 0.80±0.01, ARI = 0.22±0.04), while not giving up any risk prediction performance. A UMAP projection of the latent representations of the diagnosis sequences indeed showed that patients with the same ground truth disease label (either T1D or T2D) tended to be close in the latent space (Figure S1). VaDeSC (AUC = 0.66±0.01 and ARI = 0.11±0.09) was the next best performing method, performing better than TransVarSur_nosurv (AUC = 0.69±0.01, ARI = 0.07±0.06), in which the survival loss is turned off. TransVarSur's performance is degraded when not considering the survival times (TransVarSur_nosurv), illustrating that in achieving its superior performance, the full TransVarSur model does indeed exploit the interactions between EHR trajectories and survival times. As opposed to the simulation benchmark, performance of TransVarSur and VaDeSC on the risk prediction task was statistically indistinguishable (CI = 0.70±0.04, p-val = 0.69). Mirroring the results on the simulated data benchmark, all other models performed poorly at retrieving the ground truth clusters. Finally, RDSM again clearly stood out at the third-best performing on the risk prediction task, highlighting the importance of accounting for interactions between diagnosis sequences and survival times.

Application: TransVarSur identifies clinically and genetically relevant Crohn's disease patient subgroups

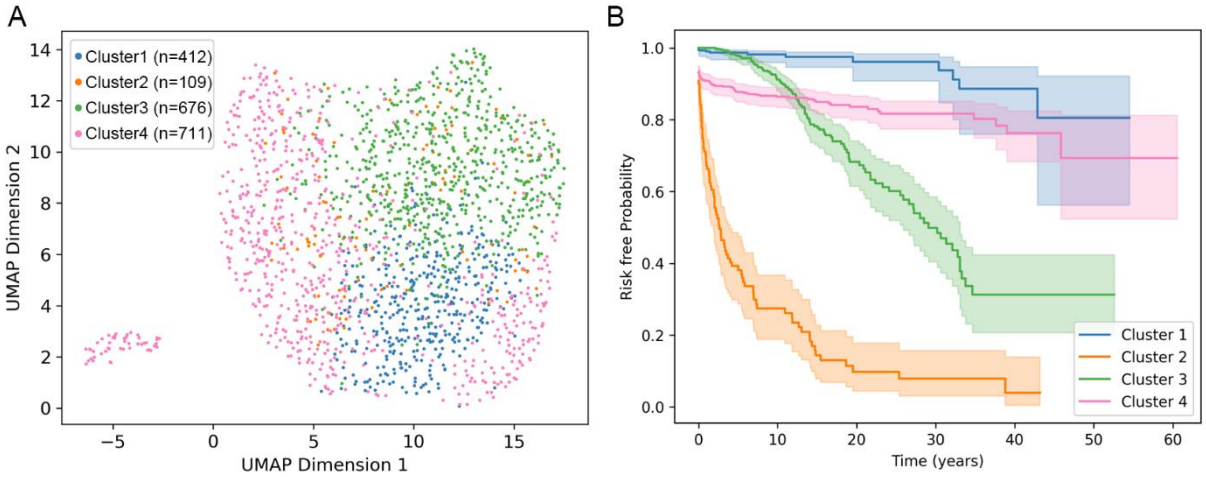

Figure 4. Clustering Crohn's disease patients from the UK Biobank in their progression towards intestinal obstruction. A) UMAP (Uniform Manifold Approximation and Projection) projection of the latent representations of the CD patients, coloring patients by cluster; B) Cluster-specific Kaplan–Meier curves with 95% confidence intervals.

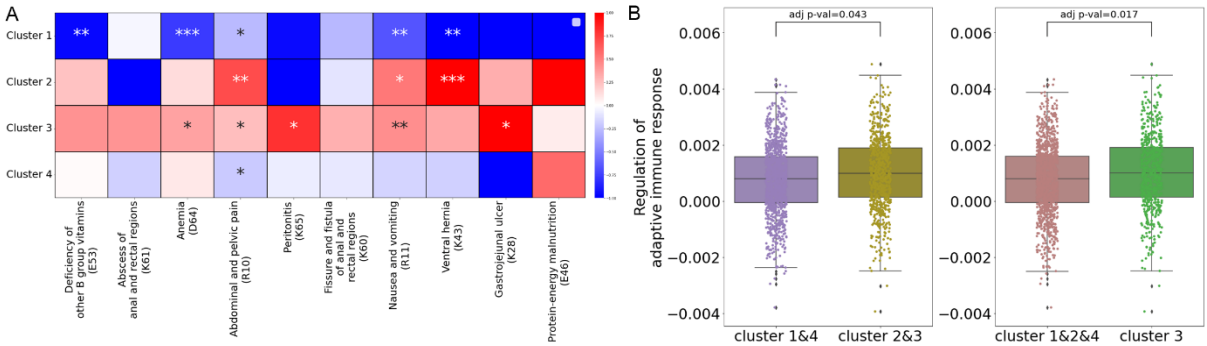

Figure 5. Association of CD clusters with individual diagnoses and pathway polygenic risk scores (pathway PRS). A) Cluster-specific enrichment of individual CD-relevant diagnoses, computed relative to all other clusters, ranging from blue (negative) to red (positive) association. Asterisks indicate the significant level: * p-val < 0.05, ** p-val < 0.01 and *** p-val < 0.001; B) Pathway PRS of the adaptive immune response pathway, comparing clusters 2 and 3 with clusters 1 and 4 (left), and cluster 3 with the other three clusters (right).

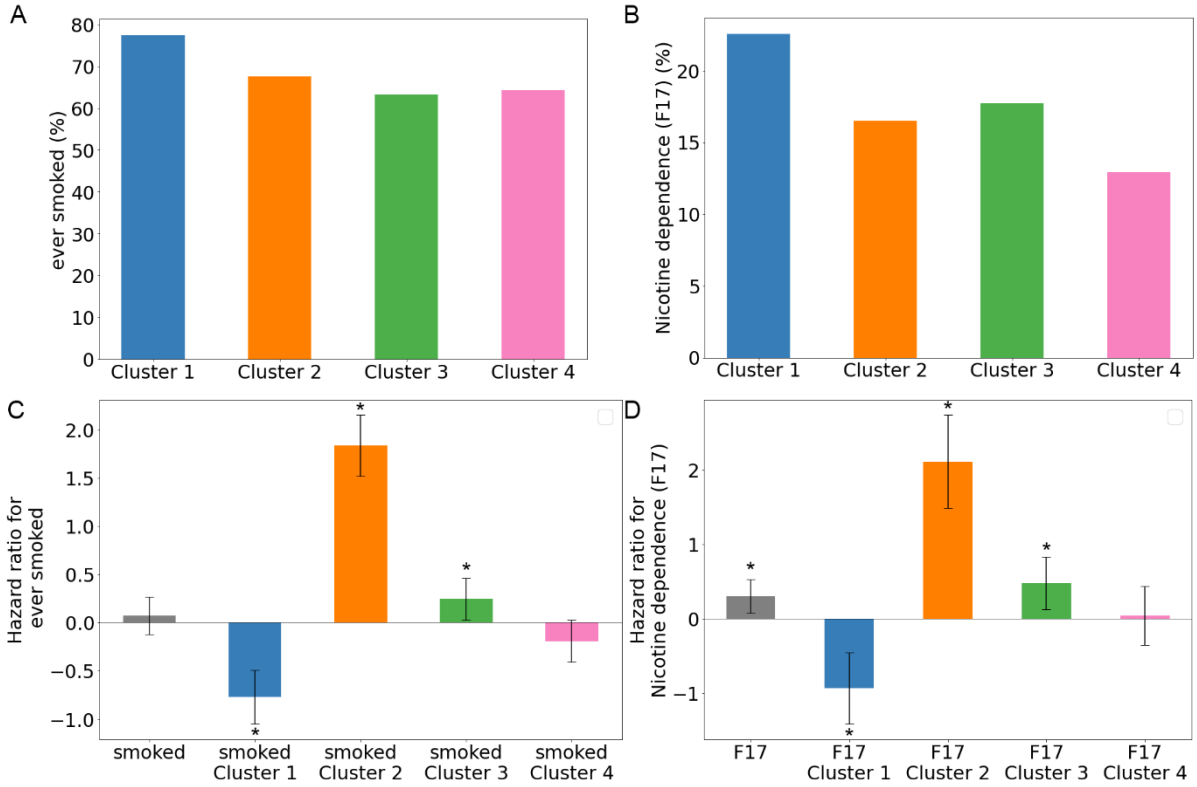

Figure 6. Association of CD clusters with smoking behavior. A) Percentage of patients who ever smoked (p-val = 1.86×10^{-05}, multinomial logistic regression with log likelihood ratio test adjusted by sex and location of location of recruitment) and B) Percentage of patients with reported nicotine dependence (ICD10 code: F17) (p-val = 0.001). C) Association of ever having smoked with risk of intestinal obstruction and D) Association of nicotine dependence with risk of intestinal obstruction (ICD10 code: F17). Error bars show the 95% confidence intervals and * represents significance of p-val < 0.05 (multivariate Cox regression).

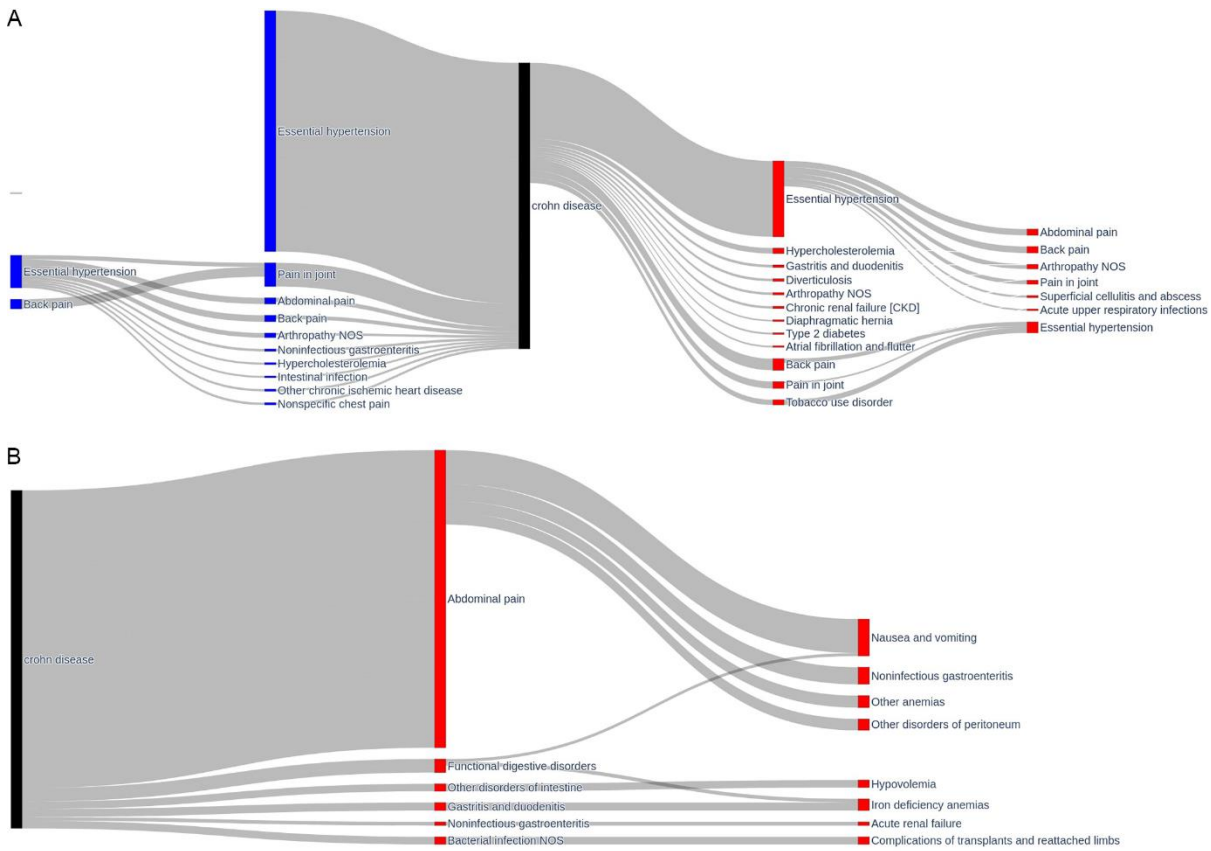

Figure 7. Enrichment of diagnosis subsequences across fast and slowly progressing clusters. Diagnoses are colored by occurrence before (blue) or after (red) the first CD diagnosis (black). A) Diagnosis subsequences significantly enriched in slow progressors (clusters 1 and 4) relative to fast progressors (clusters 2 and 3) (logistic regression coefficient > 0 and adjusted p-val < 0.05) with the width of the band representing the number of patients with a given subsequence. B) Diagnosis subsequences significantly enriched in fast progressors but not in slow progressors (logistic regression coefficient < 0 and adjusted p-val < 0.05).

After our technical validation with ground-truth cluster labels, we applied TransVarSur to 1,908 Crohn's disease (CD) patients from the UK Biobank for the purpose of identifying patient subgroups related to progression towards intestinal obstruction. Here, TransVarSur identified four patient subgroups (CI = 0.91±0.02) that demonstrated both divergent disease histories and divergent risk profiles (Figure 4). More specifically, we observed that patients clustered together tended to be close in the latent space, i.e. have diagnosis trajectories that are similar (Figure 4A). Additionally, the four patient subgroups each demonstrated distinct time-to-event profiles (Figure 4B). Analyzing variables potentially confounding our subsequent interpretation of the clusters, we found the four clusters to be significantly associated with age of onset (p-val = 6.82×10^{-05}), sex (p-val = 2.35×10^{-10}), genetic principal component number 1 (PC1) (p-val = 0.01) and UKB recruitment location (p-val = 3.91×10^{-35}), but not with the overall EHR sequence length (Figures S2 and S3). All subsequent analyses were corrected for these potential confounders.

To gain insight into the generative mechanisms giving rise to the observed differences in intestinal obstruction risk, we first analyzed differential enrichment of individual diagnoses across the clusters. We found that for the most slowly progressing patients (cluster 1), many important CD-related ICD10 codes were reported significantly less frequently, such as R10 (Abdominal and pelvic pain) (adjusted p-val = 0.03), R11 (Nausea and vomiting) (adjusted p-

val = 0.0014) and D64 (Anemia) (adjusted p-val = 0.0002) (Figure 5A). On the other hand, cluster 2 and cluster 3 with the fastest progression towards intestinal obstruction, were significantly enriched for many of the same ICD10 codes (Figure 5A). Interesting was also the association with ICD10 code K43 (Ventral hernia), significantly positively associated with clusters 2, but negatively with clusters 1 (Figure 5A). Ventral hernia is a common complication in inflammatory bowel disease [21] and a known cause of bowel obstruction [22,23], the risk event selected for the current analysis.

Due to the interesting and poorly understood differential effects of smoking behavior between the two main types of inflammatory bowel disease (protective in ulcerative colitis and harmful in Crohn's disease) [24,25], we were then interested to see whether smoking was uniformly harmful across our identified CD patient subgroups. We observed that the slowest progressing cluster 1 was enriched for smoking behavior (ever smoked: p-val = 1.86×10^{-05}; nicotine dependence: p-val = 0.001) (Figure 6A/B). Additionally, we indeed observed that smoking behavior significantly associated with faster progression towards intestinal obstruction in the overall CD patient population (nicotine dependence: hazard ratio = 0.31, p-val = 0.008) (Figure 6D). Surprisingly though, specifically in cluster 1 we observed that smoking was in fact significantly associated with slower progression towards obstruction specifically (ever smoked, hazard ratio = -0.77, p-val = 5.80×10^{-08}, Figure 6C; nicotine dependence, hazard ratio = -0.94, p-val = 0.0001, Figure 6D). Interestingly, this could suggest that in CD, patient subgroups exist for which smoking protects against progression towards intestinal obstruction, like for ulcerative colitis.

As mentioned above, patients in clusters 1 and 4 generally took longer to develop intestinal obstruction than patients in clusters 2 and 3 (Figure 4B). To further gain insight into how disease trajectories longitudinally differed between these fast and slow progressors, we assessed which diagnosis subsequences were significantly enriched between these two pairs of clusters. Although there was no significant difference in overall diagnosis sequence length between the clusters (Figure S2A), the diagnosis history leading up to the first CD diagnosis was twice as long for patients in clusters 1 and 4 (slow progressors) as it was for patients in clusters 2 and 3 (fast progressors) (Figure S4), with many diagnosis subsequences significantly enriched in the slow progressors before their first CD diagnosis relative to the fast progressors (Figure 7A). These enriched diagnosis subsequences contained many known comorbidities of CD such as back pain, hypertension, and joint pain [26]. Additionally, in clusters 1 and 4, there was no significant enrichment of serious complications after the first Crohn's disease diagnosis except hypertension. Conversely, clusters 2 and 3 were hardly significantly enriched for medical complaints leading up to their first CD diagnosis (Figure 7B). However, the patients in clusters 2 and 3 did appear to progress more rapidly as evident from developing serious complications after CD onset, including complications due to abdominal pain, nausea, vomiting and iron deficiency anemia[26-28] (Figure 7B). Interestingly, although patients in clusters 1 and 4 generally had a better prognosis than patients in clusters 2 and 3, the diagnosis trajectories between clusters 1 and 4 did seem to be quite distinct (Figure S6). More specifically, we found that patients in cluster 1 were more enriched for pain (Back pain, Abdominal pain and Pain in joint) before their first CD diagnosis compared to patients in cluster 1. Meanwhile, patients in cluster 4 were more enriched for hypertension in the whole EHR trajectory compared to patients in cluster 1 (Figure S6 and Figure S7). It could be that the patients in cluster 4 have a lower abundance of intestinal bacteria (especially *Faecalibacterium prausnitzii*), which has been shown to lead to hypertension in some CD patients [29].

In addition to the EHR data that was used for identifying the four CD patient subgroups presented in this section, UK Biobank provides a wealth of other types of data on the same

CD patients that we did not use for clustering. Using the available genetics data, we computed pathways PRSs to assess differences in the underlying genetic background between fast and slow progression clusters (Table S1). Among others, we found that the fast progressors (clusters 2 and 3) displayed a higher genetic burden in a pathway related to the adaptive immune response (Figure 5B), relative to the slow progressors (clusters 1 and 4) (Figure 5B). Specifically, we found that the patients in fastest progressing cluster 2 displayed a higher genetic burden in this pathway than the patients in all other, more slowly progressing, clusters. The involvement of innate immunity versus adaptive immunity in the pathogenesis of CD is an important topic of research [30]. Recently, several studies demonstrated the role of an abnormal adaptive immune response in the pathogenesis of CD [30-32], and the over-reactive adaptive immune response is in fact the target of current CD treatments [32]. Our results confirm the important role that the adaptive immune response may play in the pathogenesis of CD in at least a sub-population of CD patients.

DISCUSSION

Precision medicine aims to develop therapies that are targeted to specific patient subgroups based on their predicted disease risk or progression, or treatment response. Due to their scale, comprehensive nature, and multimodality, EHRs have emerged as a valuable source of data for healthcare research in general and the development of precision medicine approaches specifically [33]. An important concept in precision medicine is the identification of novel patient subgroups, i.e. patient clustering. Clustering patients with similar disease trajectories can support the identification and characterization of novel disease subtypes, the molecular characterization of which can lead to the development of more personalized therapeutics approaches.

In this paper, we introduced a novel deep probabilistic model, TransVarSur, for clustering longitudinal time-to-event data extracted from EHRs. TransVarSur exploits statistical interactions between a patient's disease history and survival time, enabling it to identify non-trivial patient subgroups characterized by both divergent disease histories and survival times. We validated TransVarSur on two benchmark experiments with known ground-truth cluster labels, showing that TransVarSur outperformed all baseline methods on simultaneously the risk prediction task and the task of retrieving the ground-truth clustering. This means that, relative to TransVarSur, the baseline methods may miss clinically relevant subgroups due to either their focus on purely risk-driven clustering or their inability to model EHRs as sequences. After the benchmark validation, we applied TransVarSur to a real-world use case: clustering Crohn's disease patients in their progression towards the risk event of intestinal obstruction. We demonstrated that the subgroups identified by TransVarSur were both clinically and biologically relevant.

TransVarSur is a first attempt at integrating EHR trajectory clustering with risk modeling and we see many opportunities for future research, a few of which we list here. First, the attention mechanism in the transformer architecture could be used to improve the interpretability of resulting clusters, e.g. by integrating an attention-based feature importance score [34]. Second, in our implementation of TransVarSur we took some limited steps in representing relations between medical concepts by embedding multiple layers of the ICD10 ontology, using the alphanumeric ICD10 codes. However, more sophisticated, and potentially more powerful, approaches have been published that could be integrated with TransVarSur. Examples include modeling the entire ICD10 ontology by representing ICD10s as combinations of their

ancestors via an attention mechanism [35] as well as generating embeddings based on ICD10 text descriptions instead of the ICD10 codes [36]. Third, additional data modalities could be included to provide a more comprehensive view on a patient's disease presentation, as well as to investigate factors potentially confounding the interpretation of the clustering, such as additional longitudinal EHR data modalities (e.g. medications, lab tests and surgical procedures) and demographics (e.g. age, sex, education) or molecular data modalities commonly available in biobanks (e.g. genomics, proteomics) [37].

In conclusion, we demonstrated that TransVarSur is highly effective at disentangling complex relationships between cluster-specific disease trajectories and survival times, as retrieved from EHR data. Hence, TransVarSur can be a powerful tool for supporting the development of precision medicine approaches by its ability to discover novel risk-associated patient subgroups.

METHODS

EHR dataset from UK Biobank

For our analyses, we use both the primary and secondary (hospital) care diagnosis records made available via the UK Biobank (UKBB) resource [38]. From the total of 451,265 patients with available EHRs, we only include those with a diagnosis sequence covering a period of at least one month and containing between 5 and 200 records, which resulted in a dataset of 352,891 patients. We then map all resulting diagnosis codes from [Read v2/3 and ICD9] to ICD10 [38]. A summary of the data can be seen the Table 4.

Table 4. Summary of the data sets used in this study

CD: Crohn's disease; T2D: Type-2 diabetes

		UKBB EHR	**Benchmark (T2D/T1D)**	**CD use case**
Source	Primary care	5,999,672	60,148	53,470
	Secondary care	3,434,668	68,628	39,179
Age	0-30	219,721	937	1920
	30-60	5,468,079	52,645	51,673
	60-	3,746,540	75,194	39,056
Sex	Male	157,065	1,459	825
	Female	195,826	865	1,083
Number of records		9,434,340	128,776	92,649
Number of patients		352,891	2,324	1,908

TransVarSur architecture

Here we describe in detail the overall architecture of TransVarSur presented in Figure 1.

In our study, every patient $p \in \{1, 2, \dots, P\}$ is defined by a three-element tuple (x_p, δ_p, t_p). Specifically, $x_p$signifies the patient's sequence of diagnoses. The censoring indicator, δ_p, is assigned 0 when the survival time of the p-th patient is censored and 1 in all other cases. The censored survival time is represented by t_p, while the survival distribution $S(t|x) = P(T > t|x)$ parameters are optimized using the maximum likelihood method. To enable clustering, an unobserved cluster assignment variable $c_p \in \{1, 2, \dots, K\}$ is also considered. Consequently, our model pursues two primary goals: 1) deduce the cluster assignment c_p for each individual patient p, and 2) establish the survival distribution based on the variables x_p and c_p.

The generative process

Here, we describe the generative process of TransVarSur. First, a cluster assignment $c \in \{1, 2, \dots, K\}$ is sampled from a categorical distribution: $p(c) = Cat(\pi)$. Then a latent embedding z is sampled from a Gaussian distribution: $p(z|c) = \mathcal{N}(\mu_c, \sigma_c^2)$. The diagnoses sequence x is generated from $p(x|z)$ which is modeled by a transformer-based decoder neural network. Finally, the survival time t is generated by $p(t|z, c)$.

Survival modeling

The survival time $p(t|z, c)$ is modeled by a Weibull distribution and adjusts for right-censoring:

$$p(t|z,c) = f(t)^{\delta} S(t|z,c)^{1-\delta} = \left[\frac{k}{\lambda_c^z}\left(\frac{t}{\lambda_c^z}\right)^{k-1} exp\left(-\left(\frac{t}{\lambda_c^z}\right)^k\right)\right]^{\delta} \left[exp\left(-\left(\frac{t}{\lambda_c^z}\right)^k\right)\right]^{1-\delta} \quad (1)$$

Where $\lambda_c^z = softplus(z^T \beta_c)$, $\beta_c \in \{\beta_1, \beta_2, \dots, \beta_K\}$; $f(t)$ is the Weibull distribution, and $S(t|z, c)$ is the survival function [17].

Evidence Lower Bound

Variables x and t are independent given z. Hence, x and c are also independent, and joint distribution $p(x, t)$ can be expressed as:

$$p(x,t) = \int_z \sum_{c=1}^{K} p(x,t,z,c) = \int_z \sum_{c=1}^{K} p(x|z)p(t|z,c)p(z|c)p(c) \quad (2)$$

Since the likelihood function in Equation 2 is intractable, we maximize the lower bound of the log marginal probability:

$$\log p(x,t) \geq \mathbb{E}_{q(z,c|x,t)} \log\left[\frac{p(x|z)p(t|z,c)p(z|c)p(c)}{q(z,c|x,t)}\right] \quad (3)$$

We approximate the probability of the latent variables z and c given the observations with a variational distribution $q(z, c|x, t) = q(z|x)q(c|z, t)$. In our model, the first term $q(z|x)$ is parameterized by a transformer-based neural network. The second term is equal to the true probability $p(c|z, t)$:

$$q(c|z,t) = p(c|z,t) = \frac{p(z,t|c)p(c)}{\sum_{c=1}^{K} p(z,t|c)p(c)} = \frac{p(t|z,c)p(z|c)p(c)}{\sum_{c=1}^{K} p(t|z,c)p(z|c)p(c)} \quad (4)$$

Thus, the evidence lower bound (ELBO) can be written as

$$L(x,t) = \mathbb{E}_{q(z|x)p(c|z,t)} \log p(x|z) + \mathbb{E}_{q(z|x)p(c|z,t)} \log p(t|z,c) - D_{KL}(q(z,c|x,t)||p(z,c)) \quad (5)$$

The ELBO can be approximated using the stochastic gradient variational Bayes (SGVB) estimator to be maximized efficiently using stochastic gradient descent. For the complete derivation, we refer the reader to previous work [17].

Transformer-based encoder and decoder

Transformer neural networks are used in both the encoder and the decoder. We use the classical transformer architecture with 6 layers, 16 attention heads, a 768-dimensional latent space, and 1280-dimensional intermediate layers. The maximum sequence length is set to 200 as mentioned before.

Summarizing with a pooling layer: SeqPool

In the original BERT language model, a token [CLS] was used mainly to summarize the information from the sequence [4]. However, EHR sequences tend to be considerably longer, for instance containing ten or more visits. Relying solely on one summarization token would inevitably result in a loss of information. As a solution, we utilize a SeqPool layer to consolidate the entire sequence of a patient into a single, comprehensive embedding for that individual [39].

SeqPool maps the output sequence using the transformation $T: \mathbb{R}^{b\times n\times d} \rightarrow \mathbb{R}^{b\times d}$.

Given:

$$X_L = f(X_0) \in \mathbb{R}^{b\times n\times d}$$

where X_L is the output of an L layer transformer encoder f, and b is the batch size, n is the sequence length, d is the total embedding dimension. X_L is fed to a linear layer $g(X_L) \in \mathbb{R}^{d\times 1}$, and softmax activation is applied to the output:

$$X'_L = softmax(g(X_L)^T) \in \mathbb{R}^{b\times 1\times n}$$

This generates an importance weighting for each input token, which is applied as follows [39]:

$$z = X'_L X_L = softmax(g(X_L)^T) \times X_L \in \mathbb{R}^{b\times 1\times d}$$

By flattening, the output $z \in \mathbb{R}^{b\times d}$ is produced which is a summarized embedding for the full patient sequence.

Pre-training and fine-tuning strategy

We pre-train the encoder part of our model on the entire UK Biobank EHR dataset (Table 1, column 1). Following the original BERT paper, we use a masked diagnosis learning strategy for pre-training. Specifically, for each patient, in the diagnosis sequence, we set an 80% probability to replace a code by [MASK], a 10% probability to replace a code by a random other code, and the remaining 10% probability to keep the code unchanged.

After pre-training the TransVarSur encoder, we fine-tune TransVarSur end-to-end on specific use cases, such as Crohn's disease (CD) patients (Table 1, column 3). Important to note here is that in the fine-tuning stage, the decoder will benefit from the pre-trained encoder, because

weights are shared between the two. The number of components of the Gaussian mixture distribution is determined by minimizing the Bayesian information criterion (BIC) and maximizing the concordance index (CI). We normalized BIC to range from 0 to 1 and then maximize $\sqrt{CI^2 + (1 - BIC_{norm})^2}$.

Simulation and benchmark data

We validate and compare TransVarSur with a range of baseline methods in two settings.

1) Simulated data

Following a previously taken approach [17], we use a pseudo transformer decoder to model and simulate the autocorrelation in diagnosis sequences (Figure S8). More specifically, let K be the number of clusters, N the number of data points, L the capped sequence length, H the dimensionality of embedding, D the size of vocabulary, J the number of latent variables, k the shape parameter of the Weibull distribution and p_{cens} the probability of censoring. Then, the data generating process can be summarized as follows:

1) Let $\pi_c = \frac{1}{K}$, for $1 \leq c \leq K$
2) Sample $c_i \sim Cat(\pi)$, for $1 \leq i \leq N$
3) Sample $\mu_{c,j} \sim unif(-10,10)$, for $1 \leq c \leq K$ and $1 \leq j \leq J$
4) Sample $z_i \sim \mathcal{N}(\mu_{c_i}, \Sigma_{c_i})$, for $1 \leq i \leq N$
5) Sample $seq_i \sim unif(0, L)$, for $1 \leq i \leq N$
6) Let $g_{res}(z) = reshape(ReLU(wz + b), L \times H)$, where $w \in \mathbb{R}^{LH \times J}$ and $b \in \mathbb{R}^{LH}$ random matrices and vectors.
7) Let $x_i = g_{res}(z_i)$, for $1 \leq i \leq N$
8) Let $g_{att}(x) = softmax\left(\frac{(w_Q x + b_Q)(w_K x + b_K)^T}{\sqrt{H}} + mask\right)(w_V x + b_V)$, where w_Q, w_K, w_V and b_Q, b_K, b_V are random matrices and vectors. Mask is based on seq_i
9) Let $x_i = g_{att}(g_{att}(g_{att}(x_i)))$, for $1 \leq i \leq N$
10) Let $g_{dec}(x) = softmax(ReLU(wx + b))$, where $w \in \mathbb{R}^{D \times H}$ and $b \in \mathbb{R}^D$ random matrices and vectors.
11) Let $x_i = argmax(g_{dec}(x_i))[1{:}seq_i]$, for $1 \leq i \leq N$
12) Sample $\beta_{c,j} \sim unif(-2.5,2.5)$, for $1 \leq c \leq K$ and $1 \leq j \leq J$
13) Sample $u_i \sim Weibull\left(softplus\left(z_i^T \beta_{c_i}\right), k\right)$, for $1 \leq i \leq N$
14) Sample $\delta_i \sim Bernoulli(1 - p_{cens})$, for $1 \leq i \leq N$
15) Let $t_i = u_i$, if $\delta_i = 1$, and sample $t_i \sim unif(0, u_i)$ otherwise, for $1 \leq i \leq N$

In our experiments, we fix $K = 3, N = 30000, J = 5, D = 1998$ (ICD10 category level vocabulary), $k = 1, p_{cens} = 0.3$, $L = 100$ and for the pseudo-attention operation, we use 3 attention layers with 10 heads in each layer. We split the generated data into three parts, one for training, one for validating and one for testing the model. We repeat this process for 5 times to arrive at five performance estimates.

2) Benchmark data

There are no standard benchmark datasets for EHR-based patient clustering. Therefore, in this study, we design a benchmark based on separating Type 1 diabetes mellitus patients from Type 2 diabetes mellitus patients in their progression towards retinal complications. We select patients using the ICD10 codes E10 for Type 1 diabetes mellitus and E11.3 for Type 2 diabetes mellitus, and then label them with the ICD10 code H36 for the risk event of retinal disorders. Finally, for each patient, we delete these three ICD10 codes (E10, E11.3 and H36) from the

input diagnoses sequence to avoid information leakage, which gives us our benchmark dataset (Table 4, column 2).

Methods comparison and metrics

We compare TransVarSur to a range of well-established baselines: variational deep survival clustering (VaDeSC), the semi-supervised clustering (SSC), survival cluster analysis (sca), deep survival machines (DSM) and recurrent neural network-based DSM (RDSM), as well as k-means and regularized Cox PH as naïve baselines. Finally, to assess the influence of the survival loss on the eventual clustering, we include TransVarSur_nosurv, in which the survival loss of TransVarSur is turned off. For all methods, hyperparameters are optimized by 5-fold cross validation. We use ICD10-based TF-IDF features as the input for all methods but RDSM and TransVarSur, which allow for directly modeling sequences of events.

We evaluate the clustering performance of models, when possible, in terms of balanced accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and area under the receiver-operating characteristic (AUC). Clustering accuracy is computed by using the Hungarian algorithm for mapping between cluster predictions and ground truth labels [40]. Statistical significance of performance difference is determined using the Mann−Whitney U test.

For the time-to-event predictions, we use the concordance index (CI) to evaluate the ability of the methods to rank patients by their event risk. Given observed survival times t_i, predicted risk scores δ_i, and censoring indicators δ_i , the concordance index is defined as

$$CI = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}_{t_j<t_i}\mathbf{1}_{\eta_j>\eta_i}\delta_j}{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}_{t_j<t_i}\delta_j}$$

Differential enrichment of diagnoses between clusters

To identify potential confounding complicating the interpretation of the clusters, we check for association with sex, age, education level, location of UKB recruitment, 4 genetic principal components and overall EHR sequence length by Chi-square test and one-way ANOVA. Only sex and location of UKB recruitment significantly associates with the clustering. We calculate differential enrichment of diagnoses between clusters in two ways: 1) for individual diagnoses, and 2) for sequences of diagnoses. For the first one, we directly consider the ICD10 code. For the second one, we first map the ICD10 codes to Phecode [41] and CALIBER codes [42], which provide a higher level of abstraction in defining diseases. For each patient, we then identify all, potentially gapped, subsequences of three diagnoses from the EHR data, with the following restrictions: 1) for duplicate diagnoses in the diagnosis sequence, we only consider the first one, 2) the subsequence should contain a diagnosis of the disease under study (CD) but not contain the selected risk event (intestinal obstruction).

We assess the statistical significance of the differential enrichment using a logistic regression predicting patient cluster from subsequence occurrence, adjusting for sex and location of recruitment. We correct the resulting p-value for multiple testing using the Benjamini–Hochberg procedure, and threshold at 0.05 for significance.

Analysis of smoking behavior

We analyze the association of smoking behavior with progression towards intestinal obstruction using a multivariate cox regression, individually testing the hazard ratio of two predictors related to smoking behavior: 1) data-field 20160 (ever smoked) from UK Biobank, and 2) diagnosis ICD10 code F17 (nicotine dependence). We correct both models for potential confounding by including sex and location of recruitment as covariates into the model.

Pathway-based polygenic risk scores (PRS)

We compute pathway-based polygenic risk scores ('pathway PRSs' henceforth) using PRSet to assess genetic differences between the patient clusters in, restricting ourselves to UK Biobank participants with European ancestry [43]. Quality control steps were performed before calculating pathway PRSs, including filtering of SNPs with genotype missingness > 0.05, minor allele frequency (MAF) < 0.01 and with Hardy-Weinberg Equilibrium (HWE) p-val < $5x10^{-8}$. We focus on 164 pathways related to Crohn's disease as retrieved from the Gene Ontology – Biological Process (GO-BP) database, selected based on a literature and keyword search ("IMMUNE") (Table S1) [44,45]. We calculate pathway PRS for each pathway using variants located in coding regions. Association between each pathway PRS and the clusters is assessed using logistic regression predicting cluster from pathway PRS, adjusting for age, sex, PC1 and recruitment location. And p-val is calculated by log likelihood ratio test. We correct the resulting p-val for multiple testing using the Benjamini–Hochberg procedure and p-val < 0.05 is used as the significance threshold.

Data and Code availability

The data and code that support the findings of this study are available from github (https://github.com/JiajunQiu/TransVarSur) and UK Biobank (www.ukbiobank.ac.uk).

Competing Interests

The authors declare no competing interests.

Acknowledgements

This research has been conducted using the UK Biobank, a major biomedical database (www.ukbiobank.ac.uk).

Ethics Statement

This research has been conducted using the UK Biobank Resource under Application Number 57952.

Contributions

Model design: J.Q., J.d.J.; method development: J.Q.; data analysis: J.Q. Y.H.; definition of CD use case: J.Q., F.L., C.W., J.S., J.d.J.; writing the manuscript: J.Q., J.d.J., A.M.E., I.B.; definition and supervision of research project: J.d.J.; All authors read, edited, and approved the article.

Reference

1 Electronic Public Health Reporting. *ONC Annu. Meet.* **Available at:https://www.healthit.gov/sites/default/files/2018-12/ElectronicPublicHealthReporting.pdf** (2018).

2 Sonal, P. & Jawanna, H. Hospitals' use of electronic health records data: 2015-2017. *ONC Data Brief* (2019).

3 Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* **4**, 86 (2021). https://doi.org:10.1038/s41746-021-00455-y

4 Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* **10**, 7155 (2020). https://doi.org:10.1038/s41598-020-62922-y

5 Rongali, S. *et al.* Learning Latent Space Representations to Predict Patient Outcomes: Model Development and Validation. *J Med Internet Res* **22**, e16374 (2020). https://doi.org:10.2196/16374

6 Alexander, N., Alexander, D. C., Barkhof, F. & Denaxas, S. Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records. *Stud Health Technol Inform* **270**, 499-503 (2020). https://doi.org:10.3233/SHTI200210

7 de Jong, J. *et al.* Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain* **144**, 1738-1750 (2021). https://doi.org:10.1093/brain/awab108

8 You, N., He, S., Wang, X., Zhu, J. & Zhang, H. Subtype classification and heterogeneous prognosis model construction in precision medicine. *Biometrics* **74**, 814-822 (2018). https://doi.org:10.1111/biom.12843

9 Kumar, M., Garand, M. & Al Khodor, S. Integrating omics for a better understanding of Inflammatory Bowel Disease: a step towards personalized medicine. *J Transl Med* **17**, 419 (2019). https://doi.org:10.1186/s12967-019-02174-1

10 Castela Forte, J. *et al.* Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering. *Sci Rep* **11**, 12109 (2021). https://doi.org:10.1038/s41598-021-91297-x

11 Bretos-Azcona, P. E., Sanchez-Iriso, E. & Cabases Hita, J. M. Tailoring integrated care services for high-risk patients with multiple chronic conditions: a risk stratification approach using cluster analysis. *BMC Health Serv Res* **20**, 806 (2020). https://doi.org:10.1186/s12913-020-05668-7

12 Tanniou, J., van der Tweel, I., Teerenstra, S. & Roes, K. C. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol* **16**, 20 (2016). https://doi.org:10.1186/s12874-016-0122-6

13 Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, E108 (2004). https://doi.org:10.1371/journal.pbio.0020108

14 Chapfuwa, P., Li, C., Mehta, N., Carin, L. & Henao, R. in *Proceedings of the ACM Conference on Health, Inference, and Learning* 60–68 (Association for Computing Machinery, Toronto, Ontario, Canada, 2020).

15 Nagpal, C., Li, X. & Dubrawski, A. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE J Biomed Health Inform* **25**, 3163-3175 (2021). https://doi.org:10.1109/JBHI.2021.3052441

16 Nagpal, C., Jeanselme, V. & Dubrawski, A. W. in *SPACA.*

17 Manduchi, L. *et al.* A Deep Variational Approach to Clustering Survival Data. (2022). https://doi.org:10.3929/ethz-b-000536597

18 Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med* **3**, 96 (2020). https://doi.org:10.1038/s41746-020-0301-z

19 de Jong, J. *et al.* Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Gigascience* **8** (2019). https://doi.org:10.1093/gigascience/giz134
20 Higgins, I. *et al.* in *International Conference on Learning Representations.*
21 Heise, D. *et al.* Incisional hernia repair by synthetic mesh prosthesis in patients with inflammatory bowel disease: a comparative analysis. *BMC Surg* **21**, 353 (2021). https://doi.org:10.1186/s12893-021-01350-9
22 Poh, B. R., Sundaramurthy, S. R. & Mirbagheri, N. Left paraduodenal hernia causing small bowel obstruction. *J Gastrointest Surg* **18**, 1377-1378 (2014). https://doi.org:10.1007/s11605-014-2517-1
23 Boudiaf, M. *et al.* Ct evaluation of small bowel obstruction. *Radiographics* **21**, 613-624 (2001). https://doi.org:10.1148/radiographics.21.3.g01ma03613
24 Lakatos, P. L., Szamosi, T. & Lakatos, L. Smoking in inflammatory bowel diseases: good, bad or ugly? *World J Gastroenterol* **13**, 6134-6139 (2007). https://doi.org:10.3748/wjg.v13.i46.6134
25 Piovani, D. *et al.* Ethnic Differences in the Smoking-related Risk of Inflammatory Bowel Disease: A Systematic Review and Meta-analysis. *J Crohns Colitis* **15**, 1658-1678 (2021). https://doi.org:10.1093/ecco-jcc/jjab047
26 Sinopoulou, V. *et al.* Interventions for the management of abdominal pain in Crohn's disease and inflammatory bowel disease. *Cochrane Database Syst Rev* **11**, CD013531 (2021). https://doi.org:10.1002/14651858.CD013531.pub2
27 Abomhya, A. *et al.* Iron Deficiency Anemia: An Overlooked Complication of Crohn's Disease. *J Hematol* **11**, 55-61 (2022). https://doi.org:10.14740/jh989
28 Cai, W., Cagan, A., He, Z. & Ananthakrishnan, A. N. A Phenome-Wide Analysis of Healthcare Costs Associated with Inflammatory Bowel Diseases. *Dig Dis Sci* **66**, 760-767 (2021). https://doi.org:10.1007/s10620-020-06329-9
29 Lykowska-Szuber, L. *et al.* What Links an Increased Cardiovascular Risk and Inflammatory Bowel Disease? A Narrative Review. *Nutrients* **13** (2021). https://doi.org:10.3390/nu13082661
30 Dai, C., Jiang, M. & Sun, M. J. Innate immunity and adaptive immunity in Crohn's disease. *Ann Transl Med* **3**, 34 (2015). https://doi.org:10.3978/j.issn.2305-5839.2015.01.02
31 Geremia, A., Biancheri, P., Allan, P., Corazza, G. R. & Di Sabatino, A. Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun Rev* **13**, 3-10 (2014). https://doi.org:10.1016/j.autrev.2013.06.004
32 Sutcliffe, S. *et al.* Novel Microbial-Based Immunotherapy Approach for Crohn's Disease. *Front Med (Lausanne)* **6**, 170 (2019). https://doi.org:10.3389/fmed.2019.00170
33 Abul-Husn, N. S. & Kenny, E. E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **177**, 58-69 (2019). https://doi.org:10.1016/j.cell.2019.02.039
34 Gui, N., Ge, D. & Hu, Z. in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* Article 455 (AAAI Press, Honolulu, Hawaii, USA, 2019).
35 Choi, E., Bahadori, M. T., Song, L., Stewart, W. F. & Sun, J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *KDD* **2017**, 787-795 (2017). https://doi.org:10.1145/3097983.3098126
36 Munoz-Farre, A., Rose, H. & Cakiroglu, S. A. sEHR-CE: Language modelling of structured EHR data for efficient and generalizable patient cohort expansion. *ArXiv* **abs/2211.17121** (2022).
37 Lentzen, M. *et al.* A Transformer-Based Model Trained on Large Scale Claims Data for Prediction of Severe COVID-19 Disease Progression. *IEEE Journal of Biomedical and Health Informatics* **27**, 4548-4558 (2023). https://doi.org:10.1109/JBHI.2023.3288768

38 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015). https://doi.org:10.1371/journal.pmed.1001779
39 Hassani, A. *et al. Escaping the Big Data Paradigm with Compact Transformers*. (2021).
40 Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**, 83-97 (1955). https://doi.org:https://doi.org/10.1002/nav.3800020109
41 Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110 (2013). https://doi.org:10.1038/nbt.2749
42 Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* **1**, e63-e77 (2019). https://doi.org:10.1016/S2589-7500(19)30012-3
43 Choi, S. W. *et al.* PRSet: Pathway-based polygenic risk score analyses and software. *PLoS Genet* **19**, e1010624 (2023). https://doi.org:10.1371/journal.pgen.1010624
44 Ntunzwenimana, J. C. *et al.* Functional screen of inflammatory bowel disease genes reveals key epithelial functions. *Genome Med* **13**, 181 (2021). https://doi.org:10.1186/s13073-021-00996-7
45 Michail, S., Bultron, G. & Depaolo, R. W. Genetic variants associated with Crohn's disease. *Appl Clin Genet* **6**, 25-32 (2013). https://doi.org:10.2147/TACG.S33966

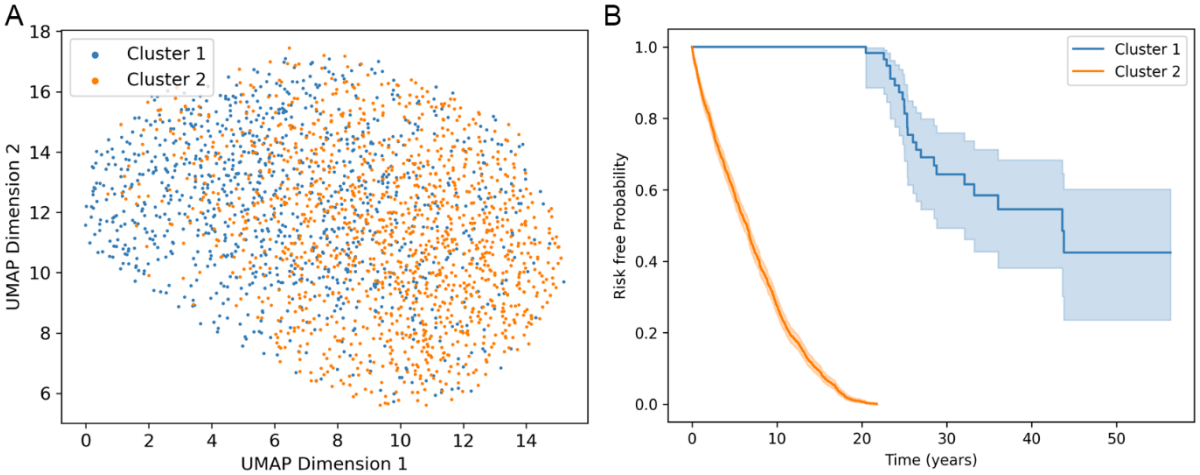

Figure S1. Using TransVarSur to distinguish T1D from T2D patients. A) UMAP of the patient representations; B) Cluster-specific Kaplan–Meier curves for both clusters.

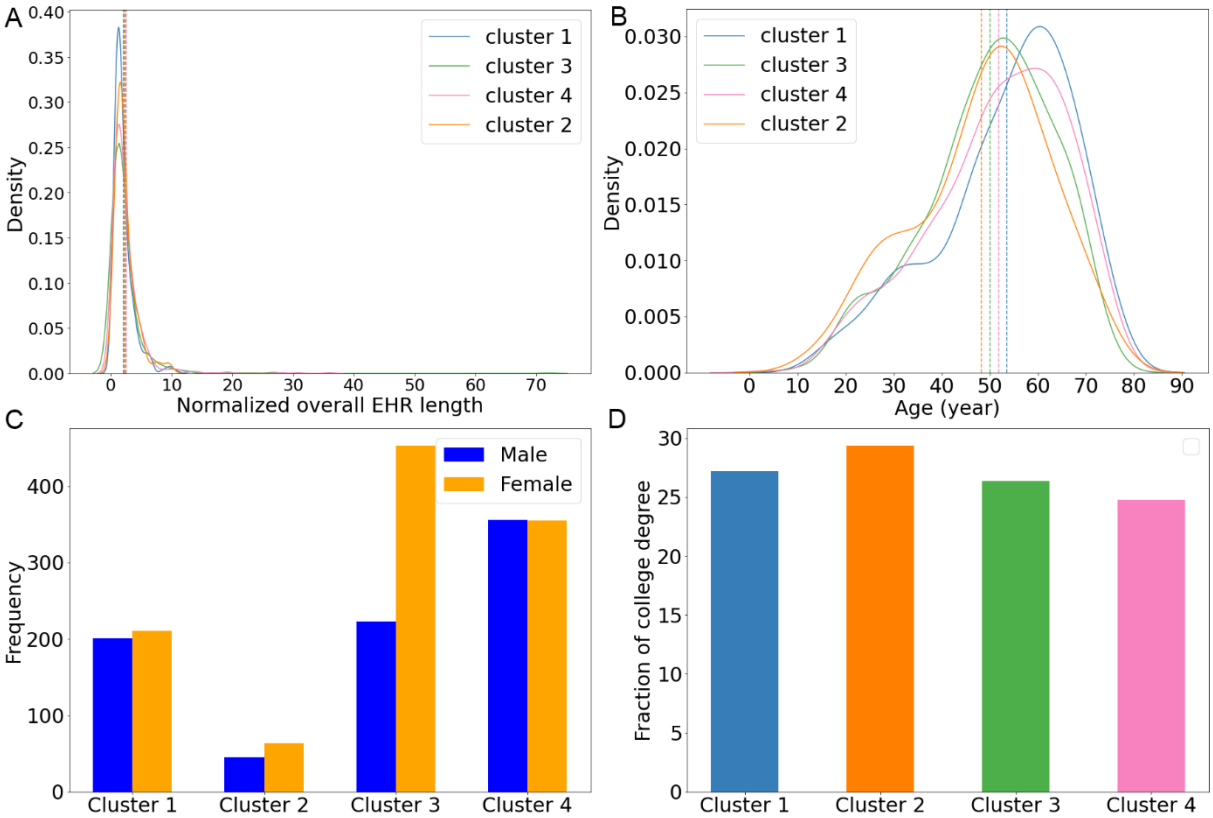

Figure S2. Association of CD patient subgroups with typical confounders. A) Normalized overall EHR length across clusters (One-way ANOVA p-val = 5.49×10^{-10}); B) Distribution of age of onset in each cluster (One-way ANOVA p-val = 6.82×10^{-05}); C) Sex distribution in each cluster (Chi-squared p-val = 2.35×10^{-10})·; D) Fraction of patients in each cluster with college degree (Chi-squared p-val =0.67).

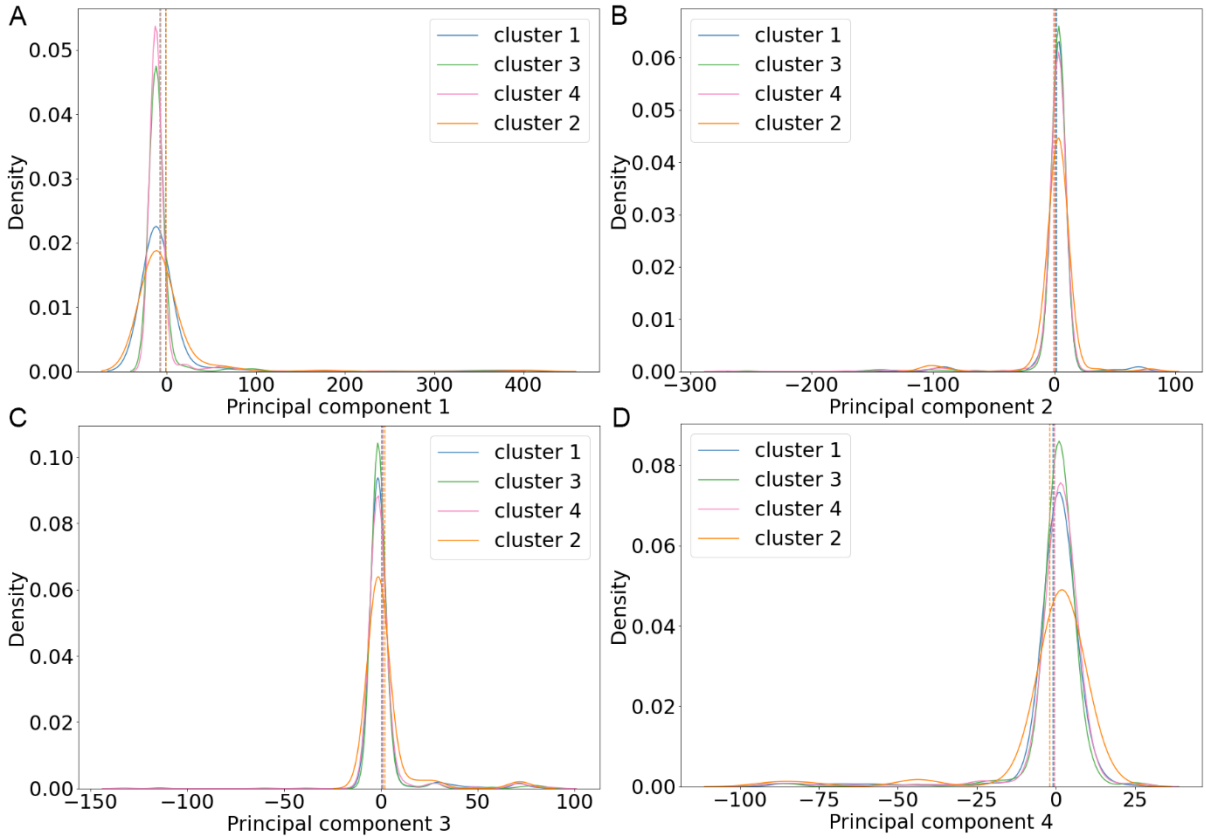

Figure S3. Distribution of genetic principal components (PCs) across the four CD patient subgroups. One-way ANOVA test was conducted to check whether there was a significant difference between the clusters (p-val = 0.01, p-val = 0.29, p-val =0.56, p-val = 0.64 for PC 1, 2, 3 and 4 respectively).

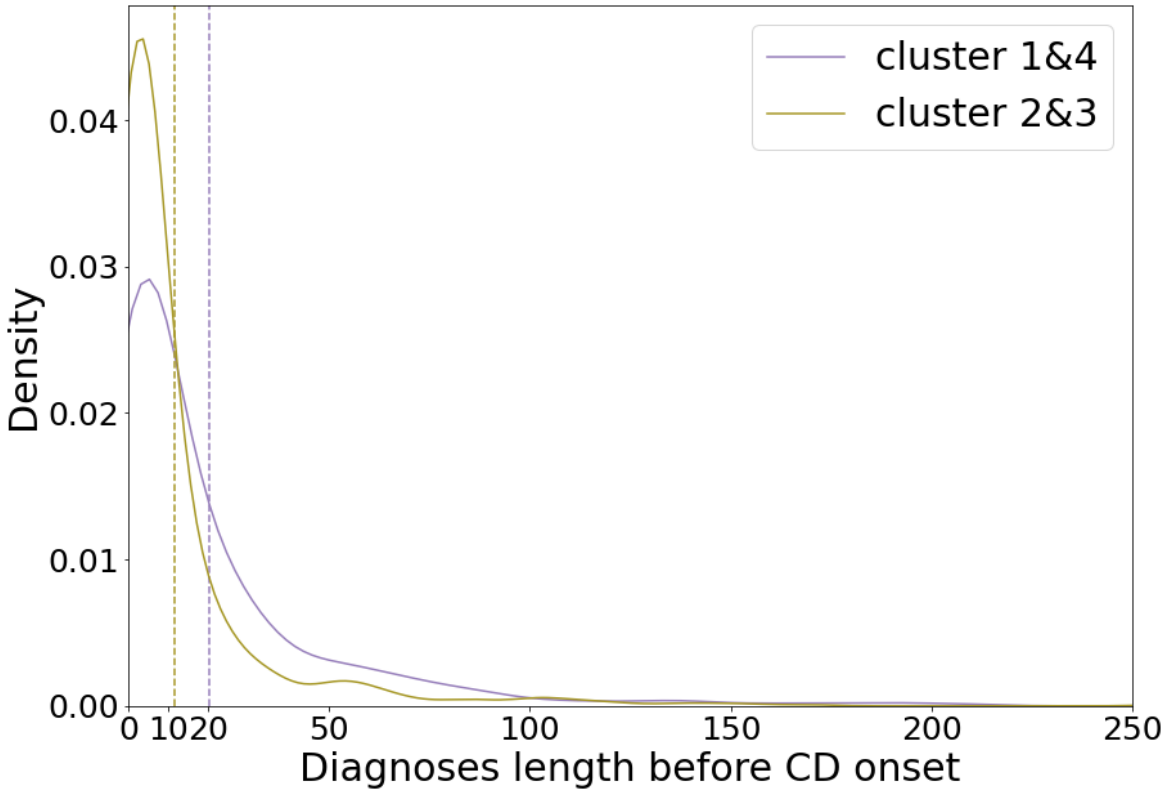

Figure S4. Distribution across clusters of EHR sequence length before first CD diagnosis. One-way ANOVA p-val = 4.48×10^{-10}.

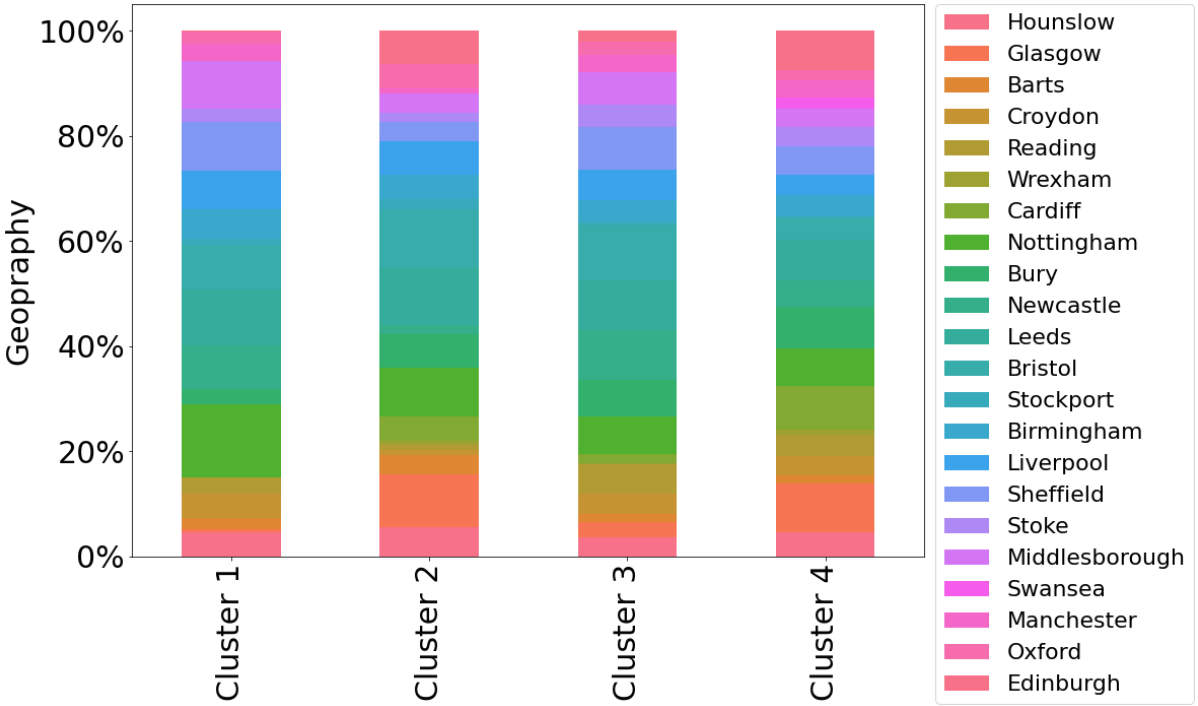

Figure S5. Association of CD patient subgroups with the location of UKB recruitment. **(**Chi-squared p-val = 3.91×10^{-35}).

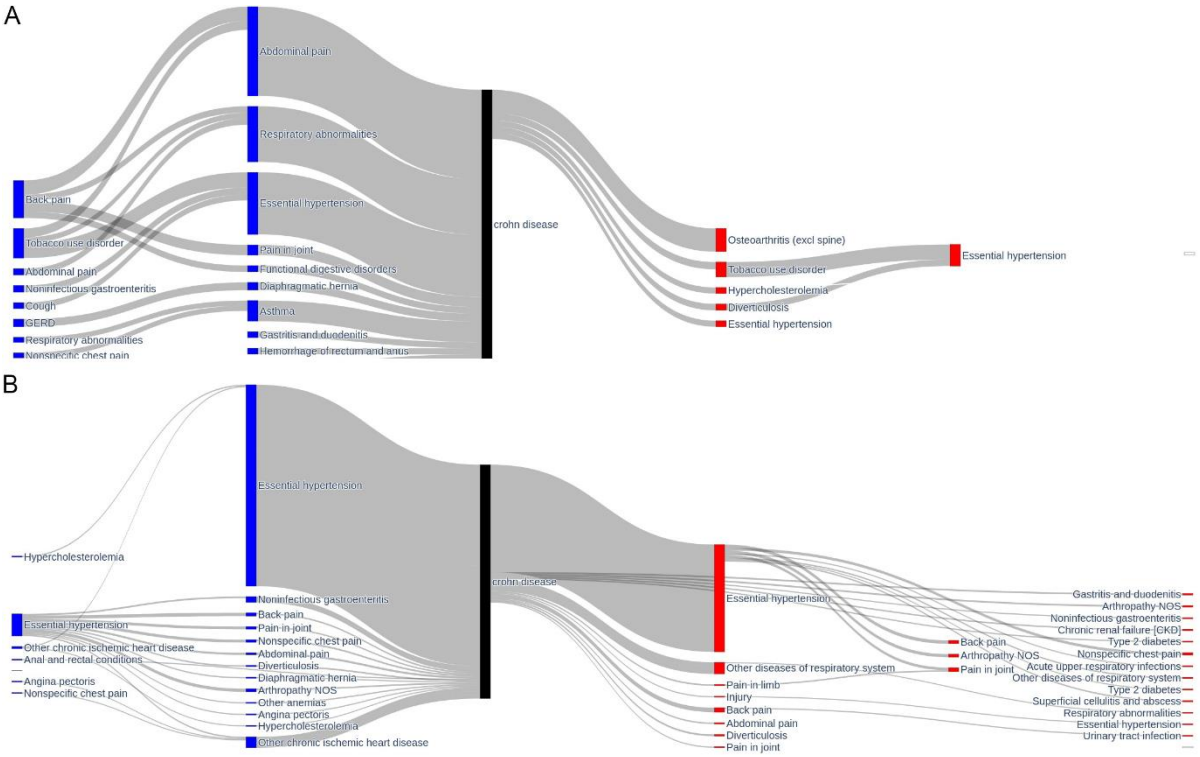

Figure S6. Association of CD clusters with diagnosis subsequences. Diagnosis subsequences enriched in cluster 1 (A) relative to cluster 4 (B) (logistic regression with a significance threshold at p-val = 0.05).

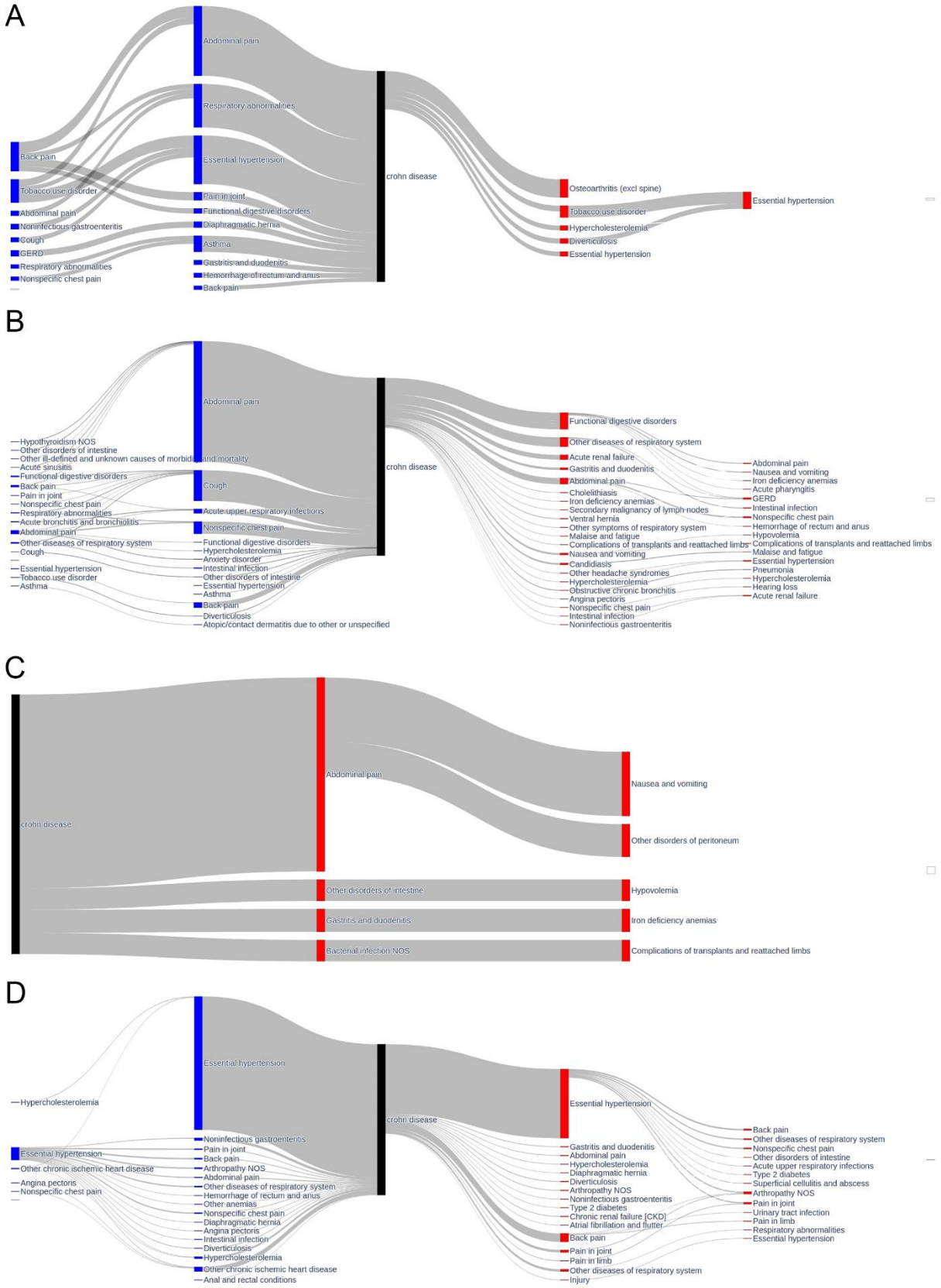

Figure S7. Differentially enriched diagnosis subsequences per cluster. A) cluster 1, B) cluster 2, C) cluster 3 and D) cluster 4. Associations were calculated by logistic regression with a significance threshold at p-val = 0.05.

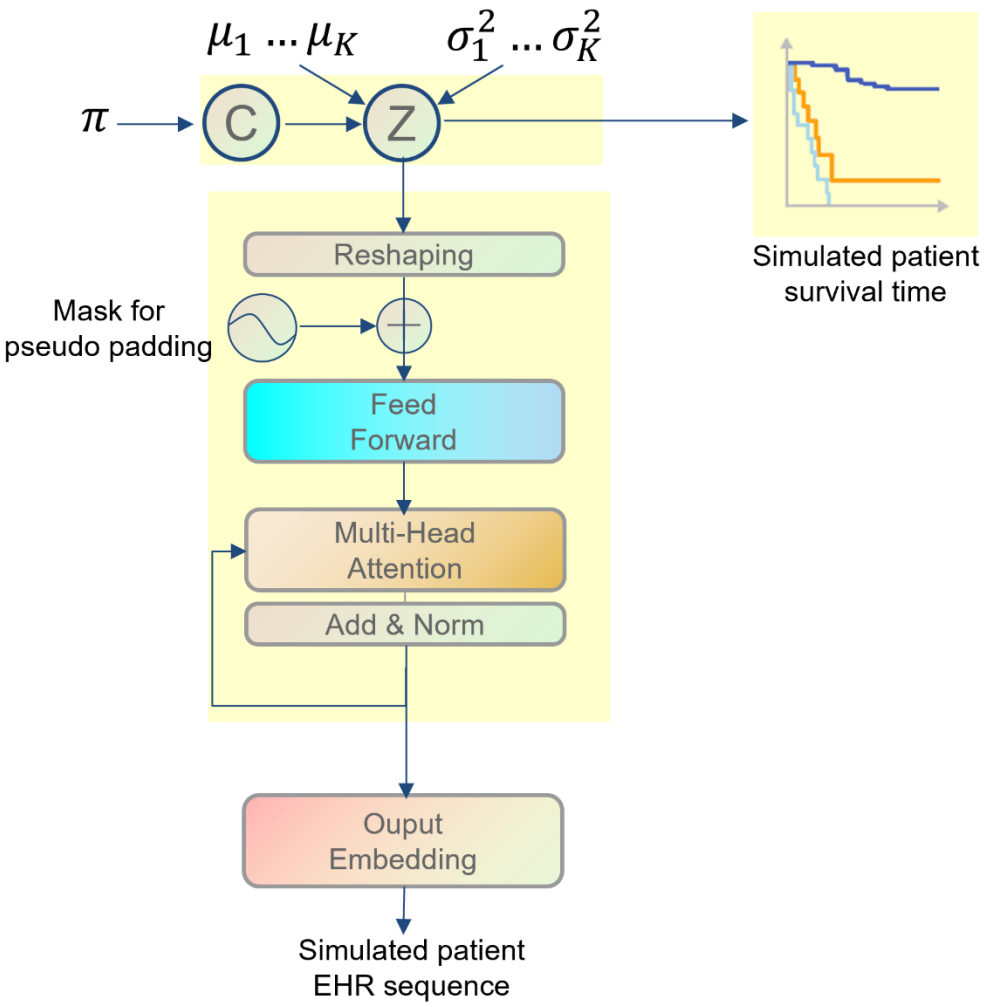

Figure S8. The generation process of simulation benchmark data. For randomly generating one EHR sequence with corresponding time-to-event, we first sample a vector Z from a Gaussian mixture distribution initialized with random parameters. Z is then used to generate both a time-to-event and an EHR sequence. The time-to-event is generated by sampling from a $Weibull$ distribution The EHR sequence is generated by feeding Z into a transformer-based pseudo-decoder with randomly initialized weights and converting to be one of ICD10 code in vocabulary with softmax function. For details we refer the reader to the Methods section.