Mixed Type Multimorbidity Variational Autoencoder: A Deep Generative Model for Multimorbidity Analysis

Woojung Kim WOOJUNG.KIM@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, West Midlands, United Kingdom

Paul A. Jenkins P.JENKINS@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, West Midlands, United Kingdom

Christopher Yau CHRISTOPHER.YAU@WRH.OX.AC.UK
Nuffield Department of Women's and Reproductive Health
University of Oxford
Oxford, Oxfordshire, United Kingdom

Abstract

This paper introduces the Mixed Type Multimorbidity Variational Autoencoder (M^3VAE), a deep probabilistic generative model developed for supervised dimensionality reduction in the context of multimorbidity analysis. The model is designed to overcome the limitations of purely supervised or unsupervised approaches in this field. M^3VAE focuses on identifying latent representations of mixed-type health-related attributes essential for predicting patient survival outcomes. It integrates datasets with multiple modalities (by which we mean data of multiple types), encompassing health measurements, demographic details, and (potentially censored) survival outcomes. A key feature of M^3VAE is its ability to reconstruct latent representations that exhibit clustering patterns, thereby revealing important patterns in disease co-occurrence. This functionality provides insights for understanding and predicting health outcomes. The efficacy of M^3VAE has been demonstrated through experiments with both synthetic and real-world electronic health record data, showing its capability in identifying interpretable morbidity groupings related to future survival outcomes.

1. Introduction

Multimorbidity refers to the acquisition of multiple long-term chronic health conditions in a single person. This is becoming an increasing public health issue with aging populations, and insights into patterns of multimorbidity are essential for managing increased health system burdens. While some health conditions may co-occur coincidentally, others exhibit a non-random correlation due to shared genetic or environmental factors. As a result, multimorbidity is no longer perceived as a random collection of individual conditions, but rather as predictable and evolving groups of conditions within individuals. This awareness has led to a growing interest in utilizing large-scale population datasets to gather evidence regarding recurring patterns of multimorbidity.

Individual-level health datasets typically consist of the health conditions that each individual possesses (binary), (continuous) physiological (e.g. blood pressure, body mass index) or blood measurements (e.g. white blood cell count), personal and demographic information (e.g. age, sex), and clinical outcome information (e.g. survival or time to event information). This data may be available as longitudinal time series if extracted from electronic health records, or in cross-sectional form if collected via surveys. While extensive methods have been developed in the context of temporal data, in this paper we focus on cross-sectional data.

Previous studies in multimorbidity clustering have employed diverse unsupervised clustering methods, such as K-means (Violán et al., 2019), Hierarchical Clustering Analysis (Roso-Llorach et al., 2018), Latent Class Analysis (Larsen et al., 2017; Hall et al., 2018; Zhu et al., 2020), Markov Clustering (Planell-Morell et al., 2020), Non-negative Matrix Factorization (Hassaine et al., 2020), and Variational Autoencoders (Gadd et al., 2022) to group patients into distinct (latent) multimorbidity clusters. Feature allocation approaches have also been seen a potential approach for identifying multimorbidity clusters where these are defined as a probability distribution over the space of morbidities (Ruiz et al., 2014; Ni et al., 2020; Kim et al., 2022; Jiang et al., 2023). These studies focus solely on clustering the binary matrices that indicate if an individual has a particular condition and do not include the other possible measurements and information available.

Alternatively, survival regression encompasses a range of models that aim to predict the survival outcomes of individuals based on input features. Traditional methods in this domain include the accelerated failure time model (Kleinbaum and Klein, 1996) and the Cox proportional hazards model (Cox, 1972). Recent advancements extend these methodologies by incorporating neural networks into the Cox method (Faraggi and Simon, 1995; Katzman et al., 2018; Kvamme et al., 2019), or directly modeling the survival distribution with neural networks (Rindt et al., 2022; Danks and Yau, 2022). These models can enable the use of mixed-type data but only provide predictive information and do not reveal any latent multimorbidity-linked structure.

Contributions. We introduce a novel deep probabilistic generative model called the Mixed Type Multimorbidity Variational Autoencoder (M^3VAE) that is designed for supervised dimensionality reduction and addresses the limitations of having purely supervised or unsupervised approaches to multimorbidity analysis. M^3VAE identifies latent representations of mixed-type health-related attributes that are particularly relevant for a patient's future survival outcome. Its distinctive capabilities include 1) the integration of diverse data sources with multiple modalities, such as health measurements, demographic information, and (possibly censored) survival outcomes, and 2) the detection of latent multimorbidity clusters, a key feature that uncovers significant patterns in disease co-occurrence, thereby providing valuable insights for understanding and predicting health outcomes. Through experiments on both synthetic and real-world electronic health record data, we demonstrate the effectiveness of our model in identifying interpretable groups of morbidities that can be related to future survival outcomes. The code implementation can be found at the following link [1].

1. https://github.com/thysics/m3vae

Generalizable insights about machine learning in the context of healthcare Treating patients with multiple morbidities presents a complex challenge, as methods effective for a single disease may be inadequate or difficult to apply in the presence of multiple concurrent conditions. This complexity has driven the development of multimorbidity analysis, an emerging field that utilizes large-scale population datasets to identify prevalent patterns of disease co-occurrence.

The main aim of multimorbidity analysis is to detect not only the patterns of diseases occurring together but also specific combinations of morbidities that significantly impact survival outcomes. Importantly, these patterns can vary across different demographic groups, underscoring the need for personalized treatment strategies that cater to the unique health profiles of individuals. Traditional methods often fall short on this objective as they tend to focus on uni-modal datasets, either highlighting morbidity patterns without considering survival outcomes and demographic factors, or forecasting survival outcome without identifying underlying multimorbidity patterns.

Addressing this gap, this paper proposes a novel machine learning model that integrates diverse data sources, including survival outcomes, demographic details, and morbidity statuses, to unearth latent multimorbidity patterns with distinct risk profiles. Our approach excels in generating low-dimensional health summaries from high-dimensional, mixed-type health data, as demonstrated through both simulated and real-world examples. Furthermore, our model uniquely identifies clinically meaningful multimorbidity patterns that no other existing models discover, highlighting its potential to discover latent patterns of co-occurring morbidites that hold significant implications for patient care. For instance, our method could be useful in identifying patients with particular comorbidities and physical characteristics. Such findings could inform population-level interventions including the development of tailored public health strategies, and focused studies on the efficacy of various medications.

2. Methods

Our goal is to develop a model that effectively compresses high-dimensional mixed-type health-related attributes into low-dimensional latent representations, facilitating the identification of latent patterns of co-occurring morbidities with notable health implications. Our model has three core objectives: i) capturing intricate dependency patterns inherent in both discrete and continuous covariates and survival outcome; ii) accommodating heterogeneity stemming from personal background factors like sex, ethnicity, and age; and iii) providing predictive insights into future outcomes.

To achieve this, we propose a deep latent variable model, otherwise known as a variational autoencoder (VAE) (Kingma and Welling, 2013), that learns the conditional joint distribution of discrete health-related covariates ($\mathbf{x}$), continuous health-related covariates ($\mathbf{c}$), and survival time (t) given personal background information ($\mathbf{b}$) through a latent variable $\mathbf{z}$, which can be expressed as follows: $p(\mathbf{x}, \mathbf{c}, t|\mathbf{b}) = \int p(\mathbf{x}, \mathbf{c}, t|\mathbf{z})p(\mathbf{z}|\mathbf{b})d\mathbf{z}$ where $p(\mathbf{z}|\mathbf{b})$ is a (conditional) prior and $p(\mathbf{x}, \mathbf{c}, t|\mathbf{z})$ is a likelihood model. To enhance model flexibility, a neural network (called *decoder*) is used to parameterize both the prior and likelihood. We train a VAE by constructing a distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{c}, \mathbf{b})$ that approximates the (intractable) posterior distribution $p(\mathbf{z}|\mathbf{x}, \mathbf{c}, t, \mathbf{b})$. The approximate distribution is parame-

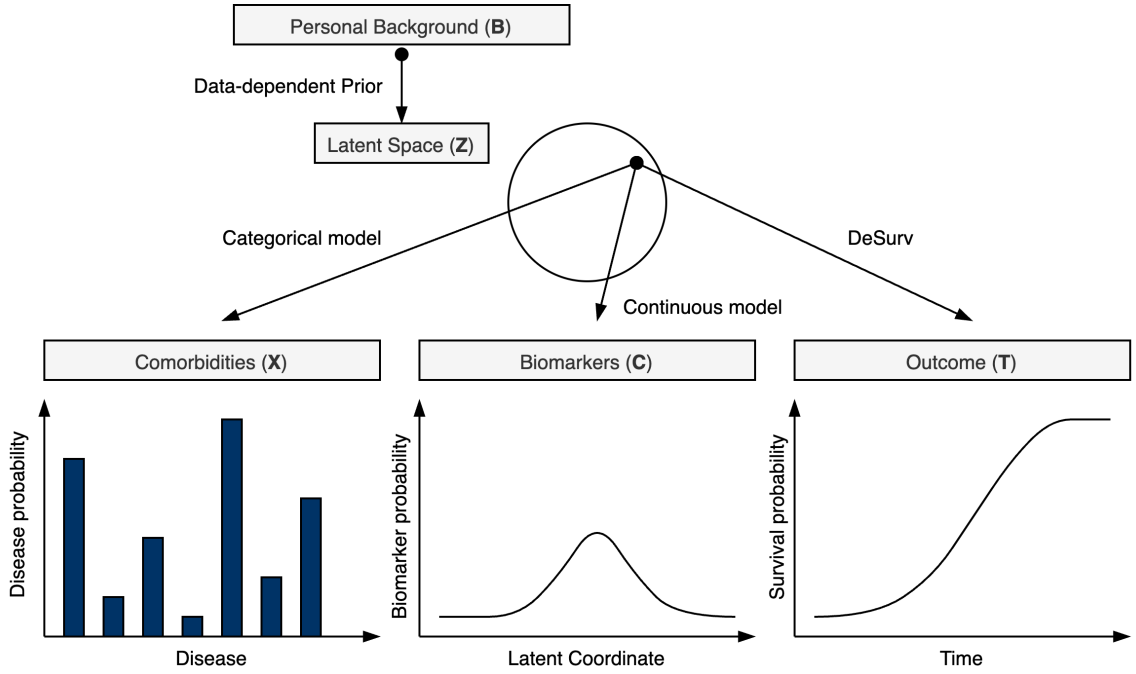

Figure 1: **Proposed Model.** The model produces individual-specific latent variables ($\mathbf{z}$) based on their personal background ($\mathbf{b}$). These latent variables are then used to simulate various health-related factors, including binary morbidities ($\mathbf{x}$), continuous biomarkers ($\mathbf{c}$), and survival time (t). The model infers a concise low-dimensional health summary represented by posterior latent variables ($\mathbf{z}$) given the observations ($\mathbf{b}, \mathbf{x}, \mathbf{c}$).

terized by a neural network called the *encoder* with input $\mathbf{x}, \mathbf{c}$, and $\mathbf{b}$. Figure 1 illustrates the generative process schematically.

Generative process In line with Khemakhem et al. (2020), we incorporate a conditional prior for the latent variables given personal background variables ($\mathbf{b}$) to account for potential variations in individuals' health conditions based on their socio-demographic characteristics. Suppose f^z is a neural network that takes $\mathbf{b}$ as its input. For any given individual i, we generate their latent variable as follows:

$$z_{il}|\mathbf{b}_i \sim \text{Laplace}(f^z_\mu(\mathbf{b}_i)_l, f^z_\sigma(\mathbf{b}_i)_l) \quad \forall l = 1, \ldots, L$$

where L is the number of latent dimensions. In other words, we employ a Laplace prior whose mean and variance are parameterized by the corresponding neural networks f^z_μ and f^z_σ, respectively. Laplace priors break the rotational invariance property of the standard isotropic Gaussian prior, therefore encouraging the learning of axis-aligned latent representations (Mathieu et al., 2019; Shi et al., 2019). This aids in achieving disentangled latent representations.

In this formulation, maximizing the data likelihood reduces the total correlation among the latent variables, as highlighted by Khemakhem et al. (2020). This promotes the emer-

gence of latent variables $(z_l)_{l=1}^L$ such that, when conditioned on personal background variables, display independence in each component (z_l). This characteristic is beneficial because it indicates the model's ability to distill high-dimensional health outcomes into lower-dimensional latent representations each of which signifies a distinct factor influencing health results, therefore facilitating "interpretable" analysis.

In our generative process, we adopt a factorized likelihood model given the latent variable, which can be expressed as: $p(\mathbf{x}, \mathbf{c}, t|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{c}|\mathbf{z})p(t|\mathbf{z})$. The distributions of covariates are modeled using standard distributions such as the Bernoulli and Gaussian distributions. To model the distribution of survival outcomes, we utilize the *DeSurv* model (Danks and Yau, 2022):

$$\begin{aligned} x_{id}|\mathbf{z}_i &\sim \text{Bernoulli}(f^x(\mathbf{z}_i)_d) \quad \forall d = 1, \ldots, D \\ c_{ij}|\mathbf{z}_i &\sim \text{Normal}(f^c(\mathbf{z}_i)_j, \sigma_j^2) \quad \forall j = 1, \ldots, J \\ \frac{du_i}{dt} &= f^u(t; \mathbf{z}_i) \\ F(t_i|\mathbf{z}_i) &= \tanh(u_i(t_i|\mathbf{z}_i)) \end{aligned}$$

where D and J correspond to the number of binary and continuous covariates, respectively. The function $F(t_i|\mathbf{z}_i)$ corresponds to the CDF of survival time which is modeled as the solution of an ordinary differential equation whose derivative is given by a positive-valued neural network f^u. This construction provides a non-parametric approach to modeling valid survival distributions as detailed in the previous section. Note that f^m denotes a neural network with input $\mathbf{z}$ which parameterises the likelihood model associated with each variable $m \in \{x, c, u\}$.

It is important to highlight that the generative process is specifically designed to enable the latent variable $(\mathbf{z})$ to function as a concise and informative health summary for each individual. Initially, from the latent health variables, we generate both covariates $(\mathbf{x}, \mathbf{c})$ and time-to-failure information (t). This enables the latent variable to act as the "fundamental" health status of an individual, influencing both diverse health measurements and their future survival outcome. Furthermore, we integrate a conditional prior for the latent variables, taking into account personal background variables $(\mathbf{b})$ to accommodate potential variations in health conditions among individuals based on their demographic characteristics. This approach further aids in identifying "interpretable" latent health summaries, where each element represents a distinct source of variation in an individual's health landscape. Consequently, this structure guarantees that the resulting latent embeddings stand as personalized and meaningful representations of an individual's health conditions.

(Variational) Inference Model training is carried out using variational inference with an approximate posterior of the form:

$$q_\phi(\mathbf{z}|\boldsymbol{\tau}) = \prod_{l=1}^{L} \text{Laplace}(z_l|g_{l;\boldsymbol{\phi}}(\boldsymbol{\tau}), \sigma^2_{l;\boldsymbol{\phi}}(\boldsymbol{\tau}))$$

where $\boldsymbol{\tau} = \{\mathbf{x}, \mathbf{c}, \mathbf{b}\}$ and L is the number of latent dimensions. In other words, a Laplace distribution is used to approximate the unknown posterior. This choice not only encourages axis-aligned representations, as demonstrated in the work of Shi et al. (2019), but also

provides a more accurate approximation of the posterior than the commonly used normal distribution in our analysis, as detailed in the Appendix A.6.

It is important to note that our inference employs an approximate posterior $q_\phi(\mathbf{z}|\boldsymbol{\tau})$ that is independent of t, therefore allowing for survival predictions at test time. This approach positions our model in a "hybrid" regime. It incorporates aspects of a "supervised" survival regression model to enable test-time predictions, while simultaneously enhancing the model's ability to perform a traditionally "unsupervised" task (i.e. dimensonality reduction): unveiling hidden data patterns that reveal population segments with variable mortality risks.

Our model's training process entails optimizing a set of parameters, $\boldsymbol{\phi}$ for the approximate posterior distribution and $\boldsymbol{\theta}$ for the generative model, to maximize the ELBO:

$$\sum_{i=1}^{N}[\alpha\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)}[\log p_{\boldsymbol{\theta}}(\mathbf{c}_i|\mathbf{z}_i)] + \gamma\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)}[\log p_{\boldsymbol{\theta}}(t_i|\mathbf{z}_i)] \\ + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)||p_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{b}_i))] \tag{1}$$

where α, γ correspond to hyperparameters that regulate the relative contribution of the continuous part of the likelihood and survival loss to the ELBO, respectively. In principle, it is possible also to up/down-weight the KL divergence term as in Higgins et al. (2017), but we found it sufficient to use weight 1.0 in our analysis.

We employ a mini-batch gradient descent method for optimization where a gradient estimator for the variational parameters $\boldsymbol{\phi}$ is derived via the Reparametrization trick (Kingma and Welling, 2013). The optimization is carried out using the Adam optimizer (Kingma and Ba, 2014) with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a batch size of 512 across all cases. The learning rate is set at 0.005, and we train for 1,000 epochs in each experiment. Our neural network architecture is uniform across all experiments, featuring a single hidden layer with 64 neurons and ReLU activation (Agarap, 2018). No dropout (Srivastava et al., 2014) or weight decay is applied.

Optimization: DeSurv Optimizing the parameters associated with generative models for discrete and continuous covariates is a straightforward task. However, optimizing the parameters related to the survival distribution, $p_{\boldsymbol{\theta}}(t|\mathbf{z})$, requires additional approximation steps for computational efficiency. To achieve this we assume non-informative censoring (Kleinbaum and Klein, 1996) which implies that survival time and censoring time are independent. This allows us to decompose the survival likelihood into two elements, based on the corresponding (failure) event indicator $s_i \in \{0, 1\}$. Specifically, we can write:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)}[\log p_{\boldsymbol{\theta}}(t_i|\mathbf{z}_i)] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\boldsymbol{\tau}_i)}\Big[\sum_{i:s_i=1}\log p_{\boldsymbol{\theta}}(t_i|\mathbf{z}_i) + \sum_{i:s_i=0}(\log(1 - F(t_i|\mathbf{z}_i))\Big].$$

In other words, when the event occurs, i.e. $s = 1$, we compute the density of the corresponding failure time using its PDF. However, if the event has not occurred during the study, and is censored, we approximate the corresponding probability using the probability of survival up to that time.

To estimate this probability we need to evaluate F; we follow Danks and Yau (2022) and use a polynomial approximation known as Gaussian-Legendre (GL) approximation

(Ambrosetti and Malchiodi, 2007), which involves approximating the integral of the neural network f^u over the time interval $[0, t]$ as a weighted sum of function values evaluated at the GL quadrature nodes. Specifically, we have

$$\begin{aligned} u(t|\mathbf{z}) &= \int_0^t f^u(v; \mathbf{z}) dv \\ &= \frac{t}{2} \int_{-1}^1 f^u \left(\frac{t}{2}(v+1); \mathbf{z} \right) dv \\ &\approx \frac{t}{2} \sum_{i=1}^n w_i f^u \left(\frac{t}{2}(v_i+1); \mathbf{z} \right) \end{aligned}$$

where w_i and v_i are the weights and nodes of the order n GL quadrature, respectively, which are computed using Legendre polynomials. The order of GL quadrature is set to be 15 throughout. With the CDF estimated, we can also compute the PDF of failure time, $p_{\boldsymbol{\theta}}(t)$, using the identity

$$p(t|\mathbf{z}) = (1 - F(t|\mathbf{z})^2) f^u(t|\mathbf{z}).$$

3. Related work

Supervised Dimensionality Reduction (SDR) methods for survival analysis, also known as survival clustering analysis, aim to identify latent clusters through projected data. Nagpal et al. (2021a,b) propose a mixture survival model that assigns each subject a discrete mixture membership within the latent representation space derived from their covariates through an encoding network. This concept aligns with Chapfuwa et al. (2020), who developed a deep generative model for survival outcomes where each subject is assigned to a distinct cluster within the latent space based on their covariates. Notably, Chapfuwa et al. (2020) employ a Dirichlet Process prior to model the distribution of latent memberships across the latent space. Lastly, Manduchi et al. (2021) adopt clustered latent variables modeled via a Gaussian mixture prior. Their approach differs from Chapfuwa et al. (2020) in that it is generative for both survival data and covariates, and utilizes probabilistic variational inference for model training.

Our work shares similarities with Manduchi et al. (2021)'s approach in generating both survival data and covariates but stands out in three distinct ways. First, it integrates both continuous and discrete covariates in the generation process. Second, we utilize a data-dependent prior to capture the potential interdependencies between covariates and survival outcomes, considering auxiliary variables. Third, the survival model is more adaptable, enhancing data fit and enabling the discovery of latent multimorbidity clusters with diverse mortality rates. These advancements allow our model to identify unique and meaningful multimorbidity clusters, which are not possible with the VaDeSC model by Manduchi et al. (2021), as is shown below.

4. Experiments

In the following, we evaluate the utility of M^{3}VAE using two simulated data sets and one real-world data set. Our experiments examine the trade-off between predictive performance

(in terms of predicting individual-level survival risk) and intepretability, which is defined here as the ability to identify sub-populations of individuals through the model's latent representations.

Setup. To evaluate prediction performance, we employ a range of conventional metrics such as the time-dependent concordance index (CI) (Antolini et al., 2005), integrated Brier score (IBS) (Graf et al., 1999), integrated negative binomial log-likelihood (NBLL) (Kvamme et al., 2019), and the right censored log-likelihood (LIK) (see details in the Appendix A.1). We note here that Rindt et al. (2022) showed that CI, IBS, NBLL are not proper scoring rules meaning that optimising against these criteria gives no guarantee of learning the true survival distribution for right-censored data as used here. The right censored log-likelihood is proven to be a proper score. We compare predictive performance against our nearest generative comparator, VaDeSC (Manduchi et al., 2021), and a pure prediction model, DeSurv (Danks and Yau, 2022) which is the same survival model incorporated into M^{3}VAE.

We also simultaneously examine whether the latent representations derived from any model demonstrate clustering patterns similar to those in the observational data. To measure this, we apply the Silverman bandwidth test (Silverman, 1981). This involves initially extracting the primary principal component from multi-dimensional latent representations and evaluating its probability of being multi-modal, i.e. its Silverman Score (SV), in line with the studies by Ahmed and Walther (2012) and Adolfsson et al. (2019) (see details in the Appendix A.1). In our model and VaDeSC, we utilised the posterior mean as the latent representations. For DeSurv, we extracted and used low-dimensional embeddings from intermediate (bottleneck) neural network layers.

Our approach is mainly compared against the leading survival clustering model VaDeSC (Manduchi et al., 2021), which is the only existing method that can identify latent representations that are related to survival outcomes and other health-related covariates. To make the comparison, we make certain adjustments to VaDeSC. For instance, to address its limitation in accommodating only one type of likelihood, we treat binary covariates as continuous variables by introducing Gaussian noise with a variance of 0.001. Additionally, we incorporate auxiliary variables into VaDeSC by considering them as an additional set of covariates.

M^{3}VAE has three hyperparameters: the number of latent dimensions, and the hyperparameters α and γ. For our simulated examples, we set the latent dimension to two and α to 0.7. In the real-world application, we opted for a latent dimension of 10 and $\alpha = 0.3$. While we fixed $\gamma = 1$ in our experiments, we show predictive performances of our model with both $\gamma \in \{1, 10\}$. To assess the predictive performance, we used ten-fold cross-validation, training on nine parts and testing on one. In this set-up, we included early stopping to prevent over-fitting, using ten percent of the training data as a validation set. The reported results are derived from the test set evaluations. All algorithms used in the experiments had modest computational demands, running within hours on an Apple M1 Pro. A detailed account of the optimization and hyperparameter choices is explored in the Appendix A.2 and A.5, respectively.

Type	Method	SV	CI	IBS	INBLL	LIK
Generative Model	M^{3}VAE ($\gamma = 1$)	**0.993 ± 0.020**	0.791 ± 0.042	0.096 ± 0.011	0.312 ± 0.031	-0.566 ± 0.053
	M^{3}VAE ($\gamma = 10$)	0.791 ± 0.175	0.825 ± 0.010	**0.061 ± 0.007**	0.194 ± 0.019	-0.132 ± 0.035
	VaDeSC	0.959 ± 0.064	0.789 ± 0.022	0.091 ± 0.008	0.298 ± 0.024	-0.708 ± 0.023
Prediction Only	DeSurv	0.898 ± 0.148	**0.827 ± 0.010**	0.061 ± 0.007	**0.193 ± 0.020**	**-0.129 ± 0.043**

Table 1: **Comparing model performance for Simulated Example A.** Best performance per metric is highlighted either in black (global) or underlined (generative model).

4.1. Synthetic example

We consider two simulated examples, each of which is created from a data generating process designed to capture the potential interactions between demographic factors and multimorbidity patterns. The data generating process of the two examples is detailed in the Appendix A.3.

Example A. Simulated example (A) features two subgroups each of which possesses comparable covariate profiles; however, they exhibit different mortality rates affected by demographic variables such as age, ethnicity, and sex. This investigation is motivated by prior research, including the study by Bots et al. (2017) which highlights the role of sex in contributing to disparities in mortality rates among individuals with common morbidities, including coronary heart disease and stroke. Each variable is generated so that each individual possesses a similar set of morbidities and covariates regardless of their background, and survival times are shaped by auxiliary variables. To ensure the former, covariates (and morbidities) are drawn from the same distribution across the population. We achieve the latter by setting the survival time as a function of two background variables b_{cont} and b_{cat} as follows: as b_{cont} increases, the survival time (t) diminishes. Separately, when $b_{cat} = 1$, it is linked to a reduced survival time. In other words, the survival distribution is dependent on background variables ($\mathbf{b}$) but independent of covariates ($\mathbf{x}, \mathbf{c}$). This implies that an effective model should concentrate exclusively on the dissimilarities in background variables rather than the covariates to identify sub-populations within a population that share the same covariates.

Table 1 shows M^{3}VAE ($\gamma = 1$) outperforms its primary competitor, VaDeSC, across all performance metrics. M^{3}VAE ($\gamma = 1$) also gives higher Silverman Score (SV) compared to the pure survival prediction methods demonstrating the utility of explicitly modelling the latent structure. As a pure prediction model, DeSurv has worse performance than both M^{3}VAE ($\gamma = 1$) and VaDeSC in terms of SV score but has better predictive performance. We next increased the weighting of the survival prediction component and found that M^{3}VAE ($\gamma = 10$) gives similar predictive performance to DeSurv. Note that in particular on the proper scoring right-censored log-likelihood measure (LIK), M^{3}VAE ($\gamma = 10$) approaches the performance of the best-performing DeSurv whose model is embedded within M^{3}VAE. This example therefore highlights the impact of the weighting factor γ which controls the relative interpretability and predictive performance of M^{3}VAE. Figure 2(A) shows that M^{3}VAE, along with VaDeSC, has the ability to detect latent clusters with different

Type	Method	SV	CI	IBS	INBLL	LIK
Generative Model	M^3VAE ($\gamma = 1$)	**0.969 ± 0.039**	0.657 ± 0.024	0.079 ± 0.009	0.254 ± 0.028	-1.716 ± 0.040
	M^3VAE ($\gamma = 10$)	0.611 ± 0.319	0.711 ± 0.019	0.069 ± 0.007	0.220 ± 0.023	-1.589 ± 0.033
	VaDeSC	0.959 ± 0.064	0.638 ± 0.021	0.087 ± 0.010	0.287 ± 0.028	-1.903 ± 0.042
Prediction only	DeSurv	0.507 ± 0.289	**0.712 ± 0.021**	**0.068 ± 0.007**	**0.215 ± 0.022**	**-1.574 ± 0.036**

Table 2: **Comparing model performance for Simulated Example B.** Best performance per metric is highlighted either in black (global) or underlined (generative model).

mortality rates. It accurately reconstructs the underlying survival functions, revealing specific survival trends unique to each hidden subgroup.

Example B. Simulated dataset (B) further supports these observations. Dataset B is composed of binary morbidity indicators ($\mathbf{x}$), continuous covariates ($\mathbf{c}$), and survival outcomes (and the event indicator), denoted as t and s, respectively. The dataset also includes personal background variables, both categorical (b_{cat}) and continuous (b_{cont}). Its data generating process includes two key characteristics: 1) The morbidity profile of each individual is contingent on their background. We model this by generating mixed-type covariates $(c, \mathbf{x})$ for each individual from a mixture distribution, with the specific distribution component they are drawn from being directed by a categorical background variable (b_{cat}). 2) The survival patterns of individuals are influenced by their covariates and a continuous background variable. To capture this, the survival times are distributed according to a Weibull distribution, the parameters of which are functions of both the covariates $(c, \mathbf{x})$ and a continuous background variable (b_{cont}). A successful model is expected to discern latent representations that reveal the underlying clusters shaped by categorical background variable as well as the continuous background variables, thereby accurately distinguishing subgroups with distinct mortality risks.

Table 2 shows M^3VAE consistently outperforms VaDeSC across all assessed metrics. However, for this more challenging dataset, DeSurv struggled to find any (implicit) representations which contain distinct clustering structure as measured by significantly lower SV scores. However, DeSurv yielded the best predictive scores. As in the previous example, increasing the weighting of the survival component in M^3VAE with $\gamma = 10$, increased predictive performance to those comparable to DeSurv but at the cost of SV decreased from 0.969 to 0.611. This further highlights the balance between interpretability and predictive performance which M^3VAE allows a user to access through γ. The discovered clustering structure is evident in Figure 2(B), where VaDeSC incorrectly identifies only two clusters instead of the actual three. This misidentification results in VaDeSC's inability to differentiate between the diverse survival trajectories associated with each cluster. On the other hand, M^3VAE successfully discerns and categorizes three distinct patterns of survival trajectories, each corresponding to the specific cluster memberships of individuals. In fact, the M^3VAE latent dimension z_0 essentially captures the risk variation.

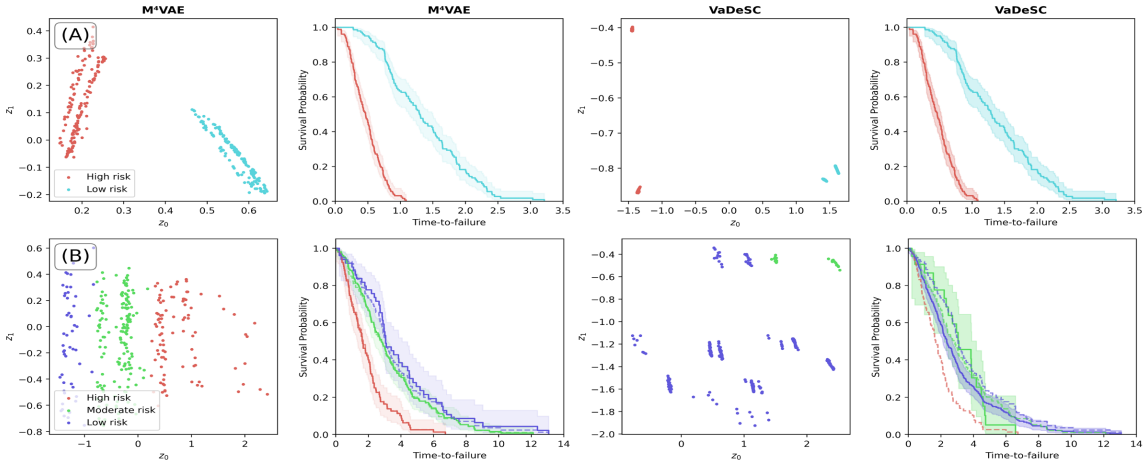

Figure 2: **Comparative analysis of simulated examples.** Two types of simulated examples: (A) Same morbidities but different mortality (B) Three distinct groups with different mortality. The left-most figure in each model, specifically M^{3}VAE and VaDeSC, represents the inferred latent space with colors indicating the respective cluster membership. The subsequent figures present estimated survival curves (bold lines) together with the corresponding ground truth (dashed lines).

4.2. Golestan study

We next evaluated the effectiveness of our model using a real-world dataset that captures a cross-sectional survey of individuals, including health conditions and background information. This dataset is collected by the study conducted in the Golestan province of Iran (Odland et al., 2021) and includes 54,000 individuals aged 36 to 81 years. Data collection occurred between 2006 and 2010, with each individual recorded only once. For the analysis, we incorporate the following variables: binary morbidity indicators ($\mathbf{x}$), continuous health measurements ($\mathbf{c}$) such as systolic/diastolic blood pressure and BMI, and socio-demographic information ($\mathbf{b}$) capturing both discrete categories (e.g. sex, ethnicity, martial status, and education level) and continuous attributes (e.g. age, wealth). After pre-processing the data, our model was trained on a dataset consisting of 11,318 patients, each with 31-dimensional features $(\mathbf{x}, \mathbf{c}, \mathbf{b})$ such that $|\mathbf{x}| = 20$, $|\mathbf{c}| = 3$ and $|\mathbf{b}| = 8$ where $|\cdot|$ signifies the cardinality. Please note that this dataset is available upon request via the NIH National Cancer Institute's GEMINI Shared Repository (GEMshare) [2]. We obtained the data by applying for it at this repository.

We assess the model's proficiency in distilling high-dimensional health attributes into lower-dimensional latent representations, which facilitates two critical outcomes: 1) robust predictions of patient survival and 2) the recognition of patient clusters with both analogous morbidity patterns and survival trajectories, thereby facilitating the identification of multimorbidity. Please note that we carried out a posterior predictive check to assess whether

2. https://dceg2.cancer.gov/gemshare/studies/GCS/

Type	Method	SV	CI	IBS	INBLL	LIK
Generative Model	M^3VAE ($\gamma = 1$)	**0.998 ± 0.003**	0.654 ± 0.031	0.104 ± 0.006	0.343 ± 0.015	-1.036 ± 0.049
	M^3VAE ($\gamma = 10$)	0.621 ± 0.296	0.724 ± 0.016	0.094 ± 0.006	0.316 ± 0.015	-0.995 ± 0.043
	VaDeSC	0.793 ± 0.342	0.514 ± 0.030	0.109 ± 0.005	0.359 ± 0.012	-1.057 ± 0.045
Prediction only	DeSurv	0.620 ± 0.237	**0.736 ± 0.013**	**0.092 ± 0.005**	**0.308 ± 0.015**	**-0.979 ± 0.047**

Table 3: **Golestan Performance Metrics.** Best performance is highlighted either in black (global) or underlined (generative model).

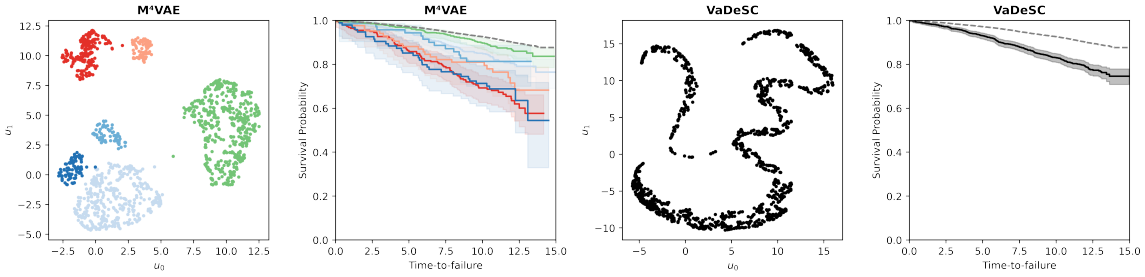

Figure 3: **Comparative analysis of Golestan study.** In each model, the left-most figure displays the inferred latent space, depicting multimorbidity clusters with distinct colors. Darker shades within each color represent higher morbidity rates. The subsequent figures show estimated survival curves for each cluster (bold lines) and the survival curve for healthy individuals (dashed lines). Notably, VaDeSC infers a single cluster.

the trained model successfully captures patterns of observed data. The outcome is present in the Appendix A.4.

Table 3 illustrates that M^3VAE is effective in creating compact latent representations of health data that are indicative of future health trajectories. It outperforms VaDeSC across all evaluated metrics. Furthermore, we replicate the phenomena seen in the simulated data and showed that the predictive performance of M^3VAE remains on par with DeSurv when the survival component is reweighted ($\gamma = 1 \rightarrow 10$) at the cost of separability in the latent space as measured by SV. Figure 3 illustrates the discovered latent clusters ($\gamma = 1$) and corresponding survival profiles which is compared to that given by VaDeSC which is unable to detect any distinct sub-populations within the entire population.

Figure 4 depicts the differentiation of six unique patient sub-groups by M^3VAE, while Table 5 in the Appendix provides a detailed breakdown of the characteristics defining each sub-group. The ability to explicitly jointly capture sub-populations and their corresponding survival profiles distinguishes M^3VAE from pure survival prediction models. For instance, clusters represented in red and light red are primarily composed of women with conditions like diabetes and dyslipidemia, consistent with prior studies that report a higher incidence of cardiometabolic comorbidities among women in Iran (Yadegar et al., 2022). Notably, the red cluster, with a larger fraction of patients suffering from stroke and higher blood pressure in comparison to the light red cluster, correlates with a reduced survival probability.

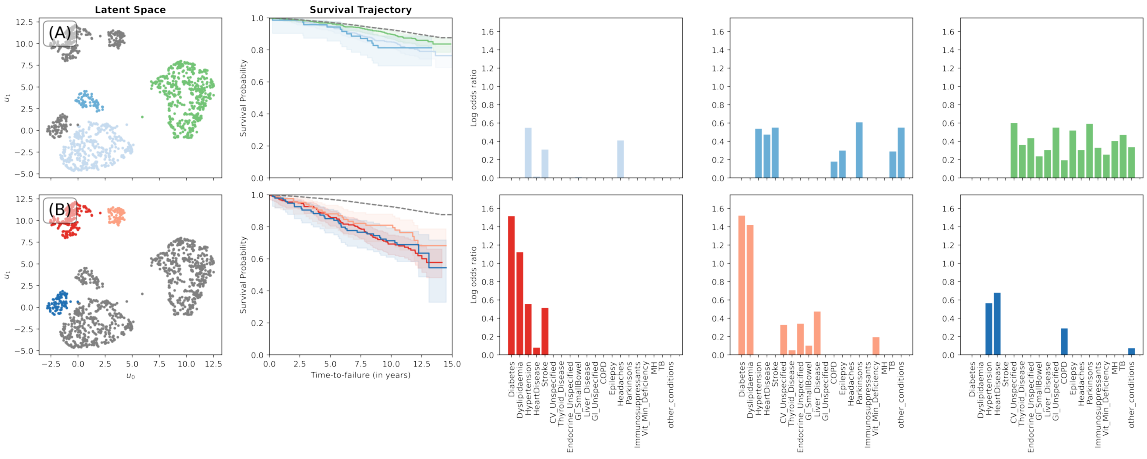

Figure 4: **Multimorbidity profiles.** The figure illustrates the inferred multimorbidity cluster, with each subfigure representing groups of multimorbidities associated with different mortality levels: (A) corresponding to low mortality rates and (B) corresponding to high mortality rates. The bar charts illustrate the log-odds ratio of the proportion of patients with each morbidity in comparison to the population.

The blue clusters primarily include hypertensive individuals. The light blue subgroup comprises individuals with high blood pressure who have suffered strokes. The sky blue subgroup consists of patients with Parkinson's disease, which is known to be associated with heart conditions. These individuals have a lower estimated survival rate compared to the general population, consistent with previous studies highlighting reduced life expectancy among Parkinson's patients with cardiovascular comorbidities (Fereshtehnejad et al., 2015). The dark blue subgroup comprises patients who have been diagnosed with both cardiovascular diseases (CVD) and chronic obstructive pulmonary disease (COPD). The higher prevalence of COPD in this subgroup can be attributed to a larger proportion of smokers compared to other groups, as indicated in Table 5. This particular cluster is associated with a significantly higher mortality rate, which aligns with previous research highlighting the increased risk of CVD-related mortality resulting from the comorbidity between CVD and COPD (Morgan et al., 2018). Lastly, the green cluster represents patients with complex and diverse multimorbidity, some of which reflects the well-established association between gastrointestinal disorders and headaches (Martami et al., 2018).

5. Discussion

We have introduced a probabilistic generative model designed for comprehensive analysis of multimorbidity. The model integrates a model of high-dimensional mixed-type health-related attributes, a low-dimensional personalized latent health representation and a survival risk model. To achieve this, the generative process encompasses three key components: 1) joint modelling between covariates and survival outcomes, 2) likelihood models for multimodal data, and 3) a data-dependent prior distribution. Through these mechanisms, we

show in our experiments that the model identifies latent morbidity clusters linked to varying risk profiles while providing survival predictions that can be comparable to those produced by state-of-the-art survival regression methods. Our model highlights and makes accessible to the user, through a single parameter, the compromise between interpretability of the latent space and predictive performance.

Limitations Our study opens several paths for future research. Currently, M^3VAE does not readily accommodate datasets comprising both structured and unstructured data within its framework. Exploring the integration of such datasets, possibly through advanced neural architectures like attention mechanisms (Vaswani et al., 2017), represents a promising direction. Additionally, adapting M^3VAE for longitudinal data analysis to track disease progression over a patient's lifetime, similar to the work of Qiu et al. (2024), offers another interesting avenue for extension.

Acknowledgments

We thank Kaspar Märtens for discussion and support in developing this work. WK is supported by the CDT in Mathematics and Statistics at the University of Warwick. PJ is supported in part by the Engineering and Physical Sciences Research Council (EPSRC) via ProbAI: A Hub for the Mathematical and Computational Foundations of Probabilistic AI (EP/Y028783/1). CY is supported by an EPSRC Turing AI Acceleration Fellowship (Grant Ref: EP/V023233/1). This work is independent research partially funded by the National Institute for Health and Care Research (NIHR, Artificial Intelligence for Multiple Long-Term Conditions (AIM), OPTIMising therapies, disease trajectories, and AI assisted clinical management for patients Living with complex multimorbidity (OPTIMAL study), NIHR202632). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health and Care Research or The Department of Health and Social Care.

References

Andreas Adolfsson, Margareta Ackerman, and Naomi C Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, 2019.

Abien Fred Agarap. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*, 2018.

Murat O Ahmed and Guenther Walther. Investigating the multimodality of multivariate data with principal curves. *Computational Statistics & Data Analysis*, 56(12):4462–4469, 2012.

Antonio Ambrosetti and Andrea Malchiodi. *Nonlinear analysis and semilinear elliptic problems*, volume 104. Cambridge university press, 2007.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Sophie H Bots, Sanne AE Peters, and Mark Woodward. Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. *BMJ global health*, 2(2):e000298, 2017.

Paidamoyo Chapfuwa, Chunyuan Li, Nikhil Mehta, Lawrence Carin, and Ricardo Henao. Survival cluster analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 60–68, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Dominic Danks and Christopher Yau. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 7240–7256. PMLR, 2022.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Seyed-Mohammad Fereshtehnejad, Azadeh Shafieesabet, Mahdiyeh Shafieesabet, Gholam Ali Shahidi, Ahmad Delbari, and Johan Lökk. Mortality in iranian patients with Parkinson's disease: Cumulative impact of cardiovascular comorbidities as one major risk factor. *Parkinson's Disease*, 2015, 2015.

Charles Gadd, Krishnarajah Nirantharakumar, and Christopher Yau. mmVAE: multimorbidity clustering using relaxed Bernoulli β-variational autoencoders. In *Machine Learning for Health*, pages 88–102. PMLR, 2022.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Marlous Hall, Tatendashe B Dondo, Andrew T Yan, Mamas A Mamas, Adam D Timmis, John E Deanfield, Tomas Jernberg, Harry Hemingway, Keith AA Fox, and Chris P Gale. Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: Latent class analysis of a nationwide population-based cohort. *PLoS medicine*, 15(3):e1002501, 2018.

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.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Xilin Jiang, Martin Jinye Zhang, Yidong Zhang, Arun Durvasula, Michael Inouye, Chris Holmes, Alkes L Price, and Gil McVean. Age-dependent topic modeling of comorbidities in UK Biobank identifies disease subtypes with differential genetic risk. *Nature Genetics*, pages 1–12, 2023.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 1–12, 2018.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Woojung Kim, Paul A Jenkins, and Christopher Yau. Feature allocation approach for multimorbidity trajectory modelling. In *Machine Learning for Health*, pages 103–119. PMLR, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

David G Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text*. Springer, 1996.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

Finn Breinholt Larsen, Marie Hauge Pedersen, Karina Friis, Charlotte Glümer, and Mathias Lasgaard. A latent class analysis of multimorbidity and the relationship to socio-demographic factors and health-related quality of life. A national population-based study of 162,283 Danish adults. *PloS one*, 12(1):e0169426, 2017.

Laura Manduchi, Ričards Marcinkevičs, Michela C Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C Neidert, Marc Pfister, et al. A deep variational approach to clustering survival data. *arXiv preprint arXiv:2106.05763*, 2021.

Fahimeh Martami, Zeinab Ghorbani, Maryam Abolhasani, Mansoureh Togha, Alipasha Meysamie, Alireza Sharifi, and Soodeh Razeghi Jahromi. Comorbidity of gastrointestinal disorders, migraine, and tension-type headache: a cross-sectional study in Iran. *Neurological Sciences*, 39:63–70, 2018.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Ann D Morgan, Rosita Zakeri, and Jennifer K Quint. Defining the relationship between COPD and CVD: what are the implications for clinical practice? *Therapeutic advances in respiratory disease*, 12:1753465817750524, 2018.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021a.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep Cox mixtures for survival regression. In *Machine Learning for Healthcare Conference*, pages 674–708. PMLR, 2021b.

Yang Ni, Peter Müller, and Yuan Ji. Bayesian double feature allocation for phenotyping with electronic health records. *Journal of the American Statistical Association*, 115(532): 1620–1634, 2020.

Maria Lisa Odland, Samiha Ismail, Sadaf G. Sepanlou, Hossein Poustchi, Alireza Sadjadi, Tom Marshall, Miles D. Witham, Reza Malekzadeh, and Justine Davies. The prevalence of multimorbidity and associations with clinical outcomes among middle aged people in Golestan, Iran: A longitudinal cohort study. *Social Science Research Network*, 2021.

Pere Planell-Morell, Madhavi Bajekal, Spiros Denaxas, Rosalind Raine, and Daniel C Alexander. Trajectories of disease accumulation using electronic health records. *Studies in health technology and informatics*, 270:469–473, 2020.

Jiajun Qiu, Yao Hu, Frank Li, Abdullah Mesut Erzurumluoglu, Ingrid Braenne, Charles Whitehurst, Jochen Schmitz, and Johann de Jong. Deep representation learning for clustering longitudinal survival data from electronic health records. *medRxiv*, pages 2024–01, 2024.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1190–1205. PMLR, 2022.

Albert Roso-Llorach, Concepción Violán, Quintí Foguet-Boreu, Teresa Rodriguez-Blanco, Mariona Pons-Vigués, Enriqueta Pujol-Ribera, and Jose Maria Valderas. Comparative analysis of methods for identifying multimorbidity patterns: a study of 'real-world' data. *BMJ open*, 8(3):e018986, 2018.

Francisco JR Ruiz, Isabel Valera, Carlos Blanco, and Fernando Perez-Cruz. Bayesian nonparametric comorbidity analysis of psychiatric disorders. *The Journal of Machine Learning Research*, 15(1):1215–1247, 2014.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Bernard W Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99, 1981.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Concepción Violán, Quintí Foguet-Boreu, Sergio Fernández-Bertolín, Marina Guisado-Clavero, Margarita Cabrera-Bean, Francesc Formiga, Jose Maria Valderas, and Albert Roso-Llorach. Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population. *BMJ open*, 9(8):e029594, 2019.

Amirhossein Yadegar, Fatemeh Mohammadi, Soghra Rabizadeh, Reihane Qahremani, Alireza Esteghamati, and Manouchehr Nakhjavani. Prevalence of different patterns of dyslipidemia in patients with type 2 diabetes in an Iranian population. *Translational Medicine Communications*, 7(1):23, 2022.

Yajing Zhu, Duncan Edwards, Jonathan Mant, Rupert A Payne, and Steven Kiddle. Characteristics, service use and mortality of clusters of multimorbid patients in England: a population-based study. *BMC medicine*, 18(1):1–11, 2020.

Appendix A. Appendix

A.1. Evaluation metrics

Concordance index The (time-dependent) concordance index (CI) (Antolini et al., 2005) serves as a metric to evaluate the accuracy of predicted risk scores for pairs of comparable patients. The underlying concept is that a "good" model should appropriately assign risk profiles to comparable pairs, distinguishing individuals who experienced death earlier from those who outlived them by assigning higher risk scores to the former.

To formally compare survival probabilities, let $\hat{F}$ represent the estimated cumulative incidence function. The concordance index evaluates the probability of the survival probability at the event time $t^{(i)}$ of an individual i being greater than the survival probability at the same event time for an individual j, given that $t^{(i)} < t^{(j)}$ and $s^{(i)} = 1$, indicating an event occurrence for individual i. Mathematically, this can be expressed as:

$$P(\hat{F}(t^{(i)}|\mathbf{x}^{(i)}) > \hat{F}(t^{(i)}|\mathbf{x}^{(j)})|t^{(i)} < t^{(j)}, s^{(i)} = 1)$$

where $\mathbf{x}^{(i)}$ corresponds to a set of covariates associated with an individual i.

In practice, the concordance index is estimated by examining the number of comparable patient pairs that the model correctly predicts in terms of their risk scores, i.e. estimated cumulative incidence function. This estimation can be calculated using the following expression:

$$\frac{\sum_{i \neq j} \mathbb{I}(s^{(i)} = 1, t^{(i)} < t^{(j)}) \mathbb{I}(\hat{F}(t^{(i)}|\mathbf{x}^{(i)}) > \hat{F}(t^{(i)}|\mathbf{x}^{(j)}))}{\sum_{i \neq j} \mathbb{I}(s^{(i)} = 1, t^{(i)} < t^{(j)})}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. For example, a time-concordance index of 0.75 signifies that the model correctly orders risk scores for a random pair with a probability of 0.75.

Brier score The Brier score (Graf et al., 1999), denoted as $\mathrm{BS}(t)$, is a metric that quantifies the mean squared error between the $\{0, 1\}$ event status at time t and the predicted cumulative incidence function $\hat{F}(t)$. To address potential bias caused by dependent censoring in right-censored data, the score is adjusted using the Inverse Probability of Censoring Weight (IPCW), assigning higher weights to subjects who remain uncensored. The survival distribution for the censoring variable, denoted by G, is often estimated by the Kaplan–Meier estimator $\hat{G}$.

The Brier score at time t is computed as follows:

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(1 - \hat{F}(t|\mathbf{x}^{(i)}))^2 \mathbb{I}(t^{(i)} < t, s^{(i)} = 1)}{\hat{G}(t^{(i)})} + \frac{\hat{F}(t|\mathbf{x}^{(i)})^2 \mathbb{I}(t^{(i)} > t)}{\hat{G}(t)} \right)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. A perfect model that accurately predicts the cumulative incidence values of 1 or 0 for individuals with and without death, respectively, would result in a Brier score of zero. On the other hand, a reference model that assigns a value of 0.5 to all patients would have a Brier score of 0.25.

To assess the Brier score across all time points, the integrated Brier score (IBS) is commonly employed. The IBS is calculated as the integral of the Brier score over the entire

test time interval $[t_1, t_2]$, defined as:

$$\text{IBS} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}(s) ds.$$

For evaluation purposes, we use the `pycox` Python package with 1000 time points.

Negative Binomial Log-likelihood Kvamme et al. (2019) propose an alternative scoring rule called the negative binomial log-likelihood, which is derived from the Brier score. Instead of using the mean square error score, they utilize the log-likelihood of the Bernoulli distribution as follows:

$$\text{NBLL}(t) = -\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\log(\hat{F}(t|\mathbf{x}^{(i)})) \mathbb{I}(t^{(i)} < t, s^{(i)} = 1)}{\hat{G}(t^{(i)})} + \frac{\log(1 - \hat{F}(t|\mathbf{x}^{(i)})) \mathbb{I}(t^{(i)} > t)}{\hat{G}(t)} \right).$$

Similar to the integrated Brier score, the negative binomial log-likelihood can also be integrated to yield a scalar-valued measure of model performance:

$$\text{INBLL} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{NBLL}(s) ds.$$

Log-likelihood The right-censored log-likelihood is defined as follows:

$$s^{(i)} \log f(t^{(i)}|\mathbf{x}^{(i)}) + (1 - s^{(i)}) \log S(t^{(i)}|\mathbf{x}^{(i)})$$

where f corresponds to the PDF of the time-to-event with $S(t) = 1 - \int_0^t f(s) ds$.

Computing this log-likelihood requires an estimate of the PDF of the time-to-event. While methods like DeSurv, SuMo-net, and M^{3}VAE allow for direct estimation of the density function, other methods do not provide this estimate. Therefore, we adopt the approach proposed by Rindt et al. (2022), which estimates the survival distribution at any given time t using their survival curve estimate. Given that each t_i denotes the point at which the survival curve jumps, we calculate the estimate as follows:

$$\hat{f}(t) = -\frac{S(t_{i+k}) - S(t_{i-k+1})}{t_{i+k} - t_{i-k+1}}$$

where $t_i < t < t_{i+1}$ with $t_0 \leq \cdots \leq t_T$. Here, k determines the width of the interval, and we set $k = 2$ throughout.

Silverman Bandwidth Test The Silverman Bandwidth Test (Silverman, 1981) is a statistical method designed to test the hypothesis that the underlying density of univariate observations $X_1, \ldots, X_n$ possesses k modes, in contrast to the alternative hypothesis of exceeding k modes. Consider the null hypothesis that the distribution of data has at most k modes. Let $\hat{f}(x; h)$ be the kernel density estimate defined by:

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h} \right).$$

Here, K is the kernel and h represents the bandwidth parameter that controls the extent of smoothing applied to the data for deriving the kernel estimate. This suggests that if

the data is strongly bimodal, for instance, a larger value of h will be required to obtain a uni-modal estimate. Consequently, the hypothesis of the density having at most k modes can be rejected if the critical bandwidth, defined by:

$$h_k = \inf\{h : \hat{f}(x; h) \text{ has at most k modes}\}$$

attains a considerable magnitude. The bootstrap method (Efron, 1992) is utilized to determine the statistical significance of the critical bandwidth. When dealing with multivariate observations, the Silverman test is applied to their primary principal component, in line with the approach of Ahmed and Walther (2012). This paper refers to the resulting p-value as the Silverman Score (SV) where we set $k = 1$ throughout.

A.2. Experimental Details

Dataset Pre-processing We standardize the continuous variables by subtracting their empirical mean and dividing by their standard deviation.

The data cleaning process for the Golestan dataset involved several key steps. Firstly, we incorporate binary morbidity indicators ($\mathbf{x}$), each of which signifies the presence/absence of morbidities such as Diabetes, Dyslipidaemia, Hypertension, Heart Disease, Stroke, Unspecified Cardiovascular disease, Thyroid disease, Unspecified endocrine disease, Gastrointestinal SmallBowel, Liver disease, Unspecified gastrointestinal disease, Chronic obstructive pulmonary disease, Epilepsy, Headaches, Parkinson's disease, Immunosuppressants, Vitamin Deficiency, Mental Health disorder, Tuberculosis, and "other conditions". Here, morbidities associated with less than 1% of the population were consolidated into a category labelled as "other conditions". Next, we only retained patients exhibiting multimorbidities, meaning individuals who had been diagnosed with a minimum of two distinct morbidities. This criterion was applied to narrow down the dataset to those cases that were more relevant for the study's objectives. Additionally, patients who had missing entries for either blood pressure or BMI were excluded from the dataset. As a result, the final dataset comprised a total of 11,318 individuals.

Clustering To determine the latent cluster membership of each subject, we employ the K-means algorithm on the latent space derived from our model. The number of latent clusters is fixed to their ground truth value for both our model and VaDeSC. In simulated examples utilizing a two-dimensional latent space, we directly apply the K-means algorithm to this space. In application to real-world examples where we set the latent dimension to exceed two, we first utilize UMAP (McInnes et al., 2018) to reduce the number of latent dimensions to two before applying K-means to the reduced space with $K = 6$. For UMAP, we utilize ten neighborhoods, cosine distance, and a minimum distance of zero.

We found that VaDeSC, although it naturally identifies latent cluster membership through a Gaussian mixture prior, exhibits instability in its clustering outcome. Specifically, the inferred probabilities of observations belonging to each cluster vary significantly based on the initial values assigned to its parameters during the training process. Furthermore, although VaDeSC is informed of the ground-truth number of clusters, which is greater than one, the model incorrectly assigns the probability of certain cluster memberships as being close to zero in many cases. To ensure a fair comparison, we train the model 50

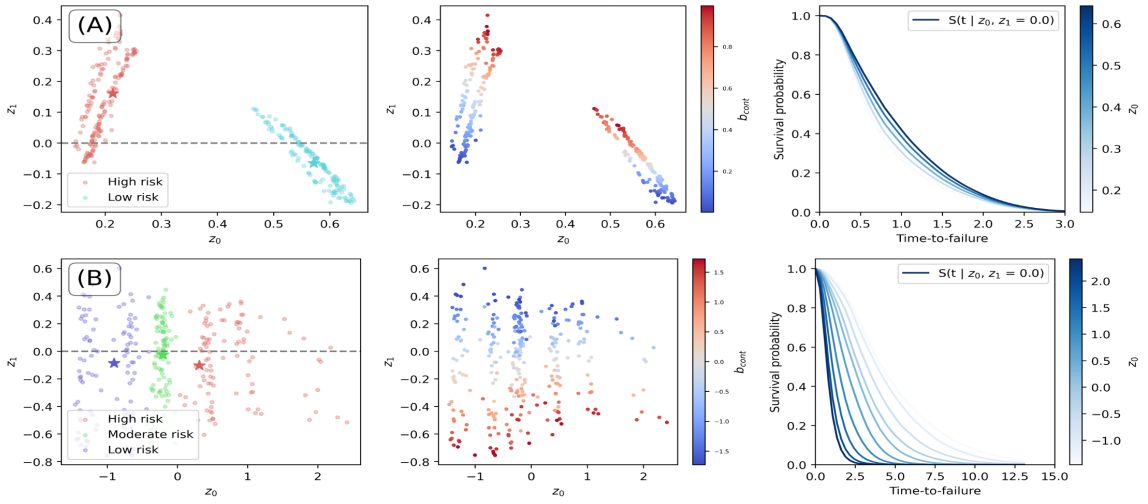

Figure 5: **Analysis of simulated data sets.** Two types of simulated examples: (A) Same morbidities but different mortality (B) Three distinct groups with different mortality. The left-most figure displays the latent space inferred by M^{3}VAE, where each point is colored according to its true cluster membership. In the middle figure, the inferred latent space is colored based on the corresponding background information b_{cont}. The right-most graph presents the predictive survival curve based on a latent dimension, while another dimension remains fixed at the value indicated by the gray vertical dashed line in the left-most figure.

Dataset	Observed	Censored	Covariates			Event time		Censoring time	
			Continuous (c)	Binary (x)	Auxiliary (b)	Mean	Max	Mean	Max
Synthetic (A)	2719 (91%)	281 (9%)	1	1	2	0.9	3.8	0.4	2.4
Synthetic (B)	2680 (89%)	320 (11%)	1	3	2	3.0	15.6	1.4	11.0
Golestan	3059 (22%)	11089 (78%)	3	20	8	6.5	14	11.7	15.0

Table 4: **Survival Datasets.** Details of the dataset used for the analysis.

times with different parameter initializations and select the outcome that yields the most probable number of clusters among these iterations.

Training All neural networks except ours and VaDeSC consist of two hidden layers with 64 units and ReLU activation (Agarap, 2018). No dropout (Srivastava et al., 2014) or weight decay is applied. The Adam optimizer (Kingma and Ba, 2014) is utilized with $\beta_1 = 0.9, \beta_2 = 0.999$, and a batch size of 512 for all cases. The implementation of VaDeSC follows their recommendation as demonstrated in Table 9 of Manduchi et al. (2021). We use a learning rate of 0.005 with 1,000 epochs for all experiments.

A.3. Synthetic Data Generation

Simulated example (A) The simulated dataset is comprised of binary morbidity indicators ($\mathbf{x}$), continuous health-related covariates ($\mathbf{c}$) and survival times (and the event

indicator), denoted as t and s, respectively, with (demographic) background variables including continuous (b_{cont}) and binary (b_{cat}) variables. We generate $N = 3,000$ individuals, with 10% experiencing (right-)censored survival times, using the data generating process below:

$$\begin{aligned}
b_{cont}^{(i)} &\sim \text{Uniform}(0,1) \\
b_{cat}^{(i)} &\sim \text{Bernoulli}\,(0.5) \\
c^{(i)} &\sim \text{Normal}(0,1) \\
x^{(i)} &\sim \text{Bernoulli}\,(0.3) \\
\lambda^{(i)} &= \exp(-(b_{cont}^{(i)} + 0.5 b_{cat}^{(i)})) \\
T^{(i)}|\lambda^{(i)} &\sim \text{Weibull}(\lambda^{(i)}, 3) \\
s^{(i)} &\sim \text{Bernoulli}\,(0.9) \\
t^{(i)}|s^{(i)} = 0 &\sim \text{Uniform}(0, T^{(i)}) \\
t^{(i)}|s^{(i)} = 1 &\sim \delta_{T^{(i)}}(t)
\end{aligned}$$

where δ_T is a Dirac mass at T and i represents the individual index.

Simulated example (B) The data generation procedure for each person i is defined as follows:

$$\begin{aligned}
b_{cont}^{(i)} &\sim \text{Uniform}(0,1) \\
b_{cat}^{(i)} &\sim \text{Categorical}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \\
x_d^{(i)}|b_{cat}^{(i)} = k &\sim \text{Bernoulli}(\rho_{kd}) \\
c^{(i)}|b_{cat}^{(i)} = k &\sim \text{Normal}(\mu_k, (0.2)^2) \\
(\lambda^{(i)}|x_d^{(i)}, c^{(i)}, a^{(i)}, b_{cat}^{(i)} = k) &= \left(1.5 - b_{cont}^{(i)}\right)\left(\sum_d x_d^{(i)} + c^{(i)}\right) \\
T^{(i)}|\lambda^{(i)} &\sim \text{Weibull}(\lambda^{(i)}, 2) \\
s^{(i)} &\sim \text{Bernoulli}\,(0.9) \\
t^{(i)}|s^{(i)} = 0 &\sim \text{Uniform}(0, T^{(i)}) \\
t^{(i)}|s^{(i)} = 1 &\sim \delta_{T^{(i)}}
\end{aligned}$$

and other hyperparameters are defined as follows:

$$\begin{aligned}
(\rho_{0d})_d &= (0.8, 0.5, 0.1), \\
(\rho_{1d})_d &= (0.8, 0.1, 0.5), \\
(\rho_{2d})_d &= (0.1, 0.5, 0.8), \\
(\mu_k)_k &= (1, 2, 3).
\end{aligned}$$

Summary statistics for key variables in both examples are presented in Table 4 and illustrations are shown in Figure 5.

	Red	Light Red	Light Blue	Sky Blue	Dark Blue	Green
# of patients	218	84	415	70	94	534
Systolic blood pressure	147.79	123.83	146.87	154.52	153.42	117.78
Diastolic blood pressure	85.96	72.54	86.60	91.65	91.03	72.90
Body mass index	30.52	26.95	28.84	27.16	28.25	26.67
Age	55.68	54.89	55.12	56.61	60.46	51.73
Sex (Male)	18%	31%	13%	27%	95%	34%
Smoker	6%	14%	5%	14%	31%	18%
Residence (City)	25%	36%	20%	29%	22%	26%
Ethnicity (Turkmens)	67%	10%	3%	13%	86%	73%
Ethnicity (Sistani)	14%	15%	1%	63%	2%	12%

Table 5: **Inferred cluster specific covariate information.** Every value represents the average of its corresponding covariate.

A.4. Golestan Study

Posterior predictive check We performed posterior predictive checks by generating data from the fitted model and comparing the baseline statistics of the generated data with the observed data. The posterior reconstruction of the data closely matches the observed data. For example, the mean of every continuous covariate and the mean of 11 out of 20 binary covariates falls within the 95% posterior interval based on 5,000 posterior simulations, as detailed below:

- **Average.DBP**: 81.041 vs. [80.160, 81.493]
- **Average.SBP**: 135.481 vs. [134.159, 136.501]
- **BMI**: 28.046 vs. [27.552, 28.143]
- **Hypertension**: 0.568 vs. [0.577, 0.619]
- **GI_SmallBowel**: 0.420 vs. [0.412, 0.462]
- **Immunosuppressants**: 0.205 vs. [0.235, 0.281]
- **MH**: 0.261 vs. [0.225, 0.267]
- **COPD**: 0.143 vs. [0.177, 0.217]

where we present the top five morbidities in frequency due to space constraints.

Additional information The demographic information for those assigned to each cluster is presented in Table 5.

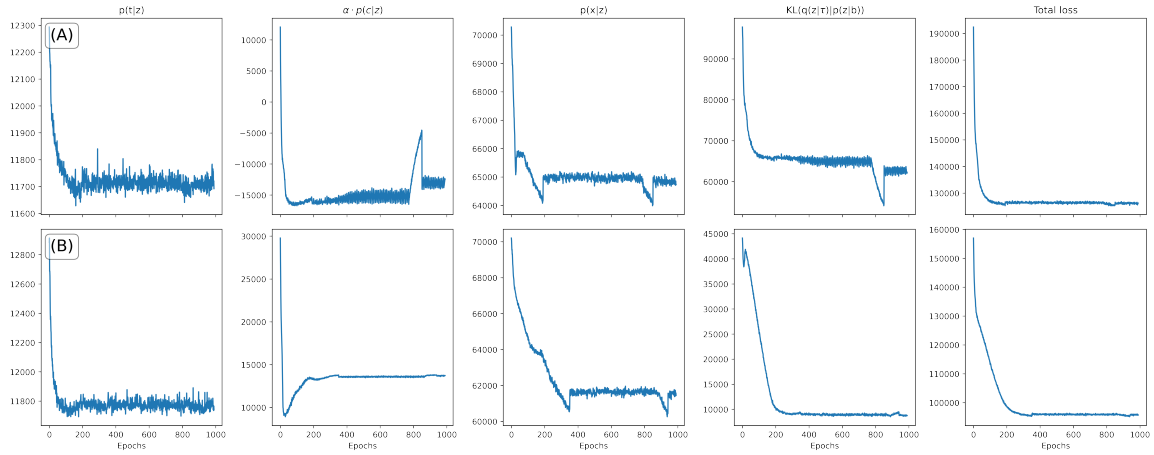

Figure 6: **Training loss trace plot: M^{3}VAE (A) $\alpha = 1.0$ (B) $\alpha = 0.3$.** The figure shows the value of each training loss term across optimization steps. The model is trained on the Golestan dataset.

A.5. Hyperparameters: α, γ

M^{3}VAE possesses hyperparameters α, γ, as shown in equation (1). The hyperparameter α determines the relative contribution of the continuous part of the likelihood to the Evidence Lower Bound (ELBO). Unlike other hyperparameters, such as γ, which allow our method to be utilized as a survival regression model, α's purpose is to address inherent scale discrepancies in the likelihood between continuous and discrete variables. By balancing these discrepancies, our model can be trained in a more balanced way with respect to each component in the loss term, thus facilitating the emergence of an interpretable latent space that reveals latent multimorbidity clustering patterns.

Figure 6 illustrates value of each loss term during the optimization process after the 10th epoch for M^{3}VAE with α set to 1.0 and 0.3. While there is a minor difference in the survival loss term across different hyperparameter specifications, $\alpha = 1.0$ results in a higher KL divergence and a lower continuous part of likelihood loss compared to $\alpha = 0.3$. This suggests that in the absence of a balancing effect controlled by α, the model might favour a local optimum of the loss where the continuous part of likelihood is better optimized at the expense of other components, such as the Bernoulli likelihood and KL divergence term. This would lead to a sparser posterior latent space, therefore hindering our ability to gain insights into latent multimorbidity clustering patterns.

The hyperparameter γ governs the relative weight of the survival loss in the ELBO, influencing the contribution of survival outcome to our model. As the value of γ increases, the model's ability to capture latent variables that closely align with the survival outcome is enhanced, leading to improved predictive performance. In Figure 7, we can observe the predictive survival curves for a random set of individuals who experienced the event in the simulated example (A). As γ increases, the resulting survival curves become more responsive to the event time. Notably, the curves exhibit a steeper decline around the event time, correctly indicating that the survival probabilities at that point are mostly below

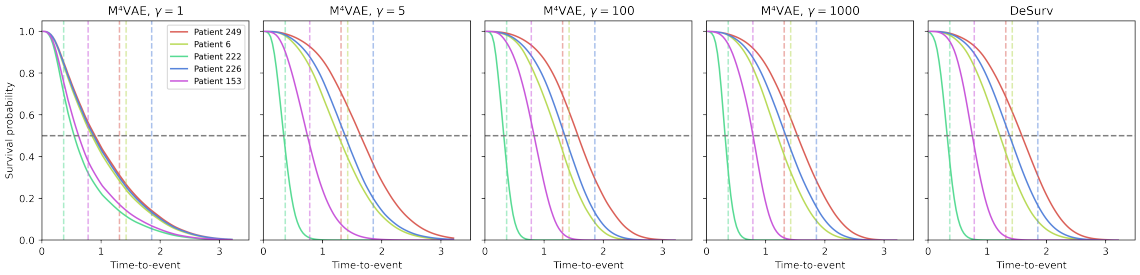

Figure 7: **Survival Curves: Simulated example (A).** The figure illustrates predicted survival trajectories for five random test patients according to each model. Each patient's survival trajectory is assigned a distinct color. Vertical dotted lines represent the event time for the corresponding patient. Horizontal grey dotted line indicates the survival probability of 0.5.

0.5. When γ is greater than or equal to 100, the resulting survival curves closely resemble those obtained from DeSurv. This finding suggests that our model can be interpreted as an extension of DeSurv, incorporating additional regularization terms that account for the mapping between latent variables and observations. Increasing the value of γ, however, can potentially compromise the accuracy of reconstructing health-related covariates from the latent variables. As a result, the resulting latent space may become less interpretable and less suitable for post-analysis interpretation.

A.6. The choice of approximate family of distributions.

Variational inference approximates the posterior by a probability distribution that is closest to the (unknown) true posterior distribution in terms of Kullback–Leibler divergence, among all distributions within a pre-chosen family. In our analysis, we use a Laplacian distribution for each latent variable to define the variational family. This decision is based on empirical evidence suggesting that the Laplace distribution provides a more accurate approximation of the (unknown) posterior distribution than the commonly used normal distribution.

To compare empirical performance of our model with different variational families, we carry out an experiment in the following steps: 1) we train our model repeatedly with different variational families (defined respectively via Normal and Laplacian distributions) with varying number of latent dimensions. In each case, 2) we compute ELBO values using the validation set. This experiment is based on the Golestan dataset.

Figure 8 demonstrates that when using a Laplacian distribution for variational inference, higher ELBO values are achieved, especially when the number of latent dimensions exceeds five. While the normal distribution does produce a slightly higher ELBO at a latent dimension that is equal to or less than five, this difference is relatively minor. Moreover, the performance disparity between the two families becomes more pronounced as the number of latent dimensions increases. These findings suggest that the unknown posterior distribution is more effectively approximated by a collection of Laplace distributions, whose parameters

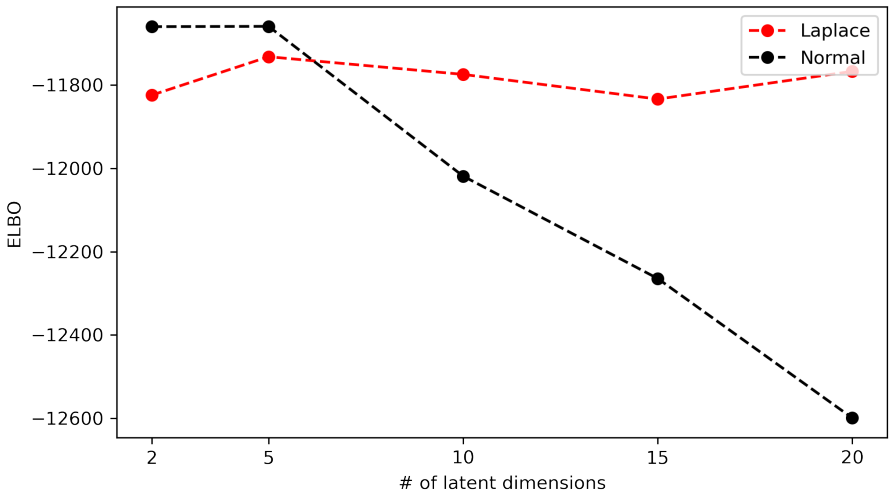

Figure 8: **Comparative analysis of two approximate distribution families.** The figure illustrates the ELBO values. It displays results for two types of approximate posterior distributions: the Normal distribution (represented in black) and the Laplace distribution (indicated in red).

are learned by a neural network, compared to the normal distribution with parameters learned in the same way.