amVAE: Age-aware multimorbidity clustering using variational autoencoders

Holm, Nikolaj Normann; Le, Thao Minh; Frølich, Anne; Andersen, Ove; Juul-Larsen, Helle Gybel; Stockmarr, Anders; Venkatesh, Svetha

Published in:
Computers in Biology and Medicine

Link to article, DOI:
10.1016/j.compbiomed.2024.109632

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Computers in Biology and Medicine 186 (2025) 109632

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

amVAE: Age-aware multimorbidity clustering using variational autoencoders

Nikolaj Normann Holm [a,*], Thao Minh Le [b], Anne Frølich [c,d], Ove Andersen [e,f,g], Helle Gybel Juul-Larsen [e], Anders Stockmarr [a], Svetha Venkatesh [b]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[b] Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia
[c] Innovation and Research Centre for Multimorbidity, Slagelse Hospital, Slagelse, Denmark
[d] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[e] Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre, Hvidovre, Denmark
[f] Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
[g] Emergency Department, Copenhagen University Hospital Amager and Hvidovre, Hvidovre, Denmark

ARTICLE INFO

Keywords:
Multimorbidity
Clustering
Disease trajectories
Variational autoencoder
Electronic health records
Chronic heart disease

ABSTRACT

Multimorbidity, the co-occurrence of multiple chronic conditions within the same individual, is increasing globally. This is a challenge for the single patients, as these individuals are subject to a heavy disease and treatment burden, yet evidence on the epidemiology and consequences of multimorbidity remains underexplored. Historically, studies aiming to understand multimorbidity patterns predominantly utilized cross-sectional data, neglecting the essential temporal dynamics which shape multimorbidity progression. Other studies based their analyses on small datasets, or populations only targeting certain sectors of the healthcare system. In this study, we (1) introduce a novel two-step multimodal Variational Autoencoder-based approach for temporal disease-based clustering (i.e. discovering age-aware multimorbidity clusters); (2) provide quantitative experiments for the robustness of our approach and the extracted temporal clusters; and (3) demonstrate how the temporal disease clusters obtained from our model can provide novel understanding of the development of multiple conditions over time and thus generate new hypotheses for different stages of multimorbidity and their associations. We trained and evaluated our models on a dataset containing the entire adult population of Denmark in the period 1995–2015, focusing on individuals suffering from chronic heart disease, including 766,596 individuals.

1. Introduction

The prevalence of multimorbidity, defined as the presence of two or more chronic conditions within the same individual [1], is on the rise globally. This increase is primarily attributed to factors such as the ageing population, due to new treatment modalities and lifestyle modifications, however counteracted by factors such as changes in physical activity levels and obesity [2,3]. Individuals suffering from multimorbidity have increased rates of mortality and hospitalizations while utilizing more bed days compared to individuals with a single condition [4]. Furthermore, as most healthcare is designed to treat individual conditions, people subject to multimorbidity often suffer a high treatment burden due to fragmented care [3,5]. Even managing the single condition becomes difficult for these individuals, as recommendations are often based on clinical trials where patients with multimorbidity were excluded [6]. As a result, there is a need for a better understanding of the underlying mechanisms related to the development of multimorbidity in order to provide appropriate, person-centered care instead of silo-based single-condition care [7].

Machine learning (ML) methodologies have emerged in multimorbidity research as an effective approach to go beyond studying conditions in isolation [8,9]. In this context, unsupervised ML methods, particularly clustering analysis, are well suited. By grouping similar patterns, clustering analysis can extract valuable insights into multimorbidity. Leveraging clinical input data, often from extensive electronic health record (EHR) databases, these methods identify groups of similar patients or diseases, facilitating a deeper understanding of the mechanisms of multimorbidity [10]. This was also the goal shared among previous cross-sectional studies [11–13]. However, while such analyses can elucidate essential disease mechanisms, they fail in capturing the

* Corresponding author.
E-mail address: nnho@dtu.dk (N.N. Holm).

https://doi.org/10.1016/j.compbiomed.2024.109632
Received 12 July 2024; Received in revised form 23 December 2024; Accepted 24 December 2024
Available online 16 January 2025

ongoing development of conditions, specifically the temporal aspects of multimorbidity. *Obtaining insights into the temporal mechanisms of multimorbidity is crucial*, as it not only facilitates the understanding of concurrent conditions, but also the time course in life when these conditions are manifested.

This importance has led to the development of clustering methodologies designed to analyze longitudinal aspects of multimorbidity from observed disease trajectories. These methodologies can broadly be categorized into three types: *disease-based clustering*, *trajectory-based clustering* and *temporal disease-based clustering*. In disease-based clustering, the aim is to gain a more comprehensive understanding of the etiology. Some of these studies adopt a network-based approach, where networks of diseases are constructed from pairs of chronic conditions. These pairs are selected based on the difference in co-occurrence prevalence compared to what would be expected based on their individual prevalences [14,15]. While the constructed networks are referred to as temporal, instead of explicitly incorporating time as a variable, temporality is based on the chronological order of conditions. Generally, these studies do not account for the significance of the time elapsed between diagnoses during the development of conditions.

On the other hand, trajectory-based studies aim to cluster patients with similar disease trajectories, thereby identifying clinically relevant patient subgroups distinguished by specific temporal disease patterns. The methodologies employed in these studies vary, ranging from non-parametric methods such as sequence analysis approaches [16] and parametric methods using the increasingly popular deep learning (DL) approaches. For instance, in Chen et al. [17], the authors trained a deep generative model using variational learning techniques [18,19] to learn a latent embedding of the entire disease trajectory. The resulting latent trajectory embeddings were subsequently clustered using the k-means algorithm. Qin et al. [20] took the idea further, incorporating predictions of future clinical outcomes into the trajectory clustering, arguing that it enhances the prognostic value of the extracted clusters. Nonetheless, the approach deviates from the realm of fully unsupervised clustering, as it introduces an additional decision point: the selection of pivotal clinical outcomes for the clustering process.

Although the discovery of patient or disease clusters based on disease trajectories is valuable, it is natural to consider disease clusters as dynamic entities present at distinct periods in a patient's life. Consequently, with a sufficiently long time horizon, it is sensible to assume that a patient can transition from one disease cluster to another throughout their lifespan. Such temporal disease-based clustering approaches, including our work, aim to identify disease clusters within distinct timeframes, with the goal of generating novel insights and hypotheses about the progression of these clusters over time. Some studies accomplish this by conducting separate clustering analyses at predetermined time points [21]. Other, more advanced methods include temporal aspects in the clustering procedure by leveraging factorization methods [22,23]. In these works, matrices or tensors are constructed, incorporating both a temporal dimension and a disease dimension in an attempt to capture the temporal dynamics of the multimorbidity progression.

In this work, we analyze the disease trajectories by decomposing them into observations of disease portfolios (unique combinations of chronic conditions) at variable-length time periods. We introduce a novel autoencoder-based clustering framework consisting of two steps. First, an unsupervised representation learning step, where we embed the multi-modal disease and age observations into a joint latent representation using a variational autoencoder. Second, a clustering step, where we cluster using the learnt latent representations in order to identify multimorbid disease clusters at different stages throughout an individual's life. We call this methodology amVAE (Age-aware Multimorbidity clustering using Variational AutoEncoders). Our code is available at https://github.com/nikolajnholm/amVAE.

For the representation learning step, we employ an autoencoder-based solution due to its effectiveness on conventional clustering benchmarks [24–26], as well as its ability to handle multiple modalities [27,28] such as disease portfolios in combination with age intervals. This AE-based approach contrasts traditional dimensionality reduction methods such as principal component analysis (PCA), a common linear preprocessing tool for clustering tasks in multimorbidity research [29–31]. Autoencoders provide a more suitable framework for our study as they can capture complex relationships among multiple data modalities each containing distinct types of features. To address the challenge of temporality, we incorporate the age interval at which disease portfolios are obtained into the feature set on which the autoencoder is trained. By treating the age modality as an interval, we implicitly model it as continuous time. The resulting joint latent space, which we call the "spatiotemporal" latent space, represents an entangled mixture of the disease portfolio and age modalities. This representation allows us to capture the interplay between disease patterns and age progression over them.

For the clustering step, we employ the k-means clustering algorithm on the learned latent representations. This choice is motivated by the advantages of hard clustering, which provides clear partitioning, simplifying the interpretability of the clusters in a clinical setting [17,20,32,33]. We apply this approach to a large dataset comprising 6,048,700 individuals, encompassing disease trajectories of the entire adult Danish population between 1995 and 2015 based on algorithmic diagnoses, incorporating information from both the primary and secondary care sectors. Specifically, we focus on individuals with chronic heart disease (HD) and aim to identify clinically meaningful clusters of conditions at different stages of the HD trajectory.

Our study provides several key contributions:

1. A novel clustering framework for temporal disease-based clustering, utilizing a multi-modal variational autoencoder.
2. Embedding continuous age-time into the feature set for temporal multimorbidity clustering.
3. Providing novel insights into the spatiotemporal structure of multimorbid disease trajectories in individuals with chronic heart disease. We uncover essential patterns within these trajectories, shedding light on the underlying dynamics of disease progression.

2. Materials and methods

2.1. Data collection and pre-processing

The data utilized for this study originates from the national Danish registers. In Denmark, every resident is legally obligated to obtain a unique personal identification number, which stores comprehensive information about the individual in the Danish Civil Registration System (CRS) [34]. Moreover, whenever a Danish resident visits a general practitioner, has contact with the hospital (acute and planned), or purchases prescribed medicine, these events are linked to their personal identification number and recorded in national registers such as the Danish National Patient Register (NPR) [35], the Danish Psychiatric Central Research Register (PCRR) [36], the Danish National Prescription Registry (DNPR) [37] and the Danish National Health Service Registry (NHSR) [38]. These registers were all utilized as data sources for this analysis. Furthermore, data on educational attainment and mortality were obtained from the Danish Population Education Register [39] and the Danish Register of causes of Death [40].

To obtain information about chronic conditions at a population level, we employed algorithmic diagnoses initially developed by the Research Center for Prevention and Health at Glostrup University Hospital, encompassing 15 chronic conditions of clinical relevance [41]. By utilizing algorithmic diagnoses, our data foundation extends beyond solely relying on ICD-10 diagnoses registered in the hospital, which predominantly cover the secondary sector. Instead, it allows us to target the population that mostly use their general practitioner for health services, resulting in a general population encompassing both the primary

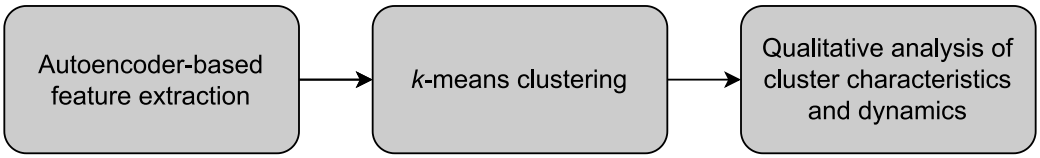


Fig. 1. Flow of the two-step approach. In the first step an autoencoder is used as feature extractor. In the second step, the extracted features are clustered using the k-means algorithm. The two-step approach results in a number of clusters which subsequently can be qualitatively analyzed.

and secondary sectors. Therefore, a specific diagnosis assigned to an individual is based on criteria related to hospitalizations, medication usage, or repeated utilization of healthcare services. Consequently, it may encompass multiple ATC and ICD-10 codes that share similar treatments and organization of healthcare (Supplementary Text A.2). Previous studies have demonstrated that the prevalence of multimorbidity, as determined by the algorithmic diagnoses, is comparable to the prevalence observed in other European countries [42]. Furthermore, these diagnoses have previously been defined and used in a longitudinal context [43].

The algorithmic diagnoses were utilized to obtain diagnostic timestamps for each of the 15 chronic conditions for all individuals 18 years or older living in Denmark from 1995–2015, comprising 6,048,700 individuals. This yielded disease trajectories for each individual, consisting of the sequence of conditions and the associated continuous timespans between them. Our study population consists of individuals subject to chronic heart disease (HD) at any point throughout the observation period, totaling 766,596 individuals whose trajectories following heart disease diagnosis have previously been studied [43].

2.2. Multimodal Autoencoder-based clustering

2.2.1. A two-step multimodal autoencoder approach for joint clustering

We propose a two-step approach for clustering the joint input of disease portfolios and age intervals. The first step consists of extracting features in the form of latent representations of the inputs through the encoder network of a trained autoencoder (AE). In the second step, clustering is performed using the latent representations and the k-means algorithm [44]. Finally, once clusters are obtained they can be analyzed qualitatively, identifying common disease portfolios and age intervals associated with the extracted clusters, as well as the dynamics related to pathways arising from cluster transitions. The approach is illustrated in Fig. 1.

The following notation is used to represent the disease trajectory data of the chronic heart disease population. Let $\mathbf{x}_p$ be a feature vector of indicators for each of the 15 considered diagnoses, i.e. a disease portfolio vector. Furthermore, let $\mathbf{x}_a$ be a feature vector with numeric information about the temporal aspects of the period the disease portfolio was observed: the starting age, stopping age and length of the timespan. Thus, a single observation related to one disease portfolio for a single individual is given by the concatenation of the two modalities, the disease portfolio and the age-interval $\mathbf{x}^{(i)} = [\mathbf{x}_p^{(i)}, \mathbf{x}_a^{(i)}]$. Naturally, as the individuals are observed over time, multiple observations are associated with each individual; these observations are treated independently of each other in the model. We aim to learn a joint compact representation of the two modalities to cluster observations based on a combination of chronic conditions and age, which facilitates the differentiation of observations based on the dynamics of an individual's medical conditions over time.

A straightforward approach to obtain such a representation is to use a vanilla AE [45], which is a neural-network-based technique that can be viewed as a non-linear extension of PCA. While PCA is limited to capturing linear relationships in the data, AEs utilize deep learning to map it to a compressed, latent representation through a series of non-linear transformations. The AE consists of two main components, an encoder and a decoder. The encoder g transforms the original input $\mathbf{x}$ into a lower-dimensional latent representation $\mathbf{z} = g(\mathbf{x})$. The decoder f, recovers the original data from the $\mathbf{z}$, creating a reconstructed input $\mathbf{x}' = f(\mathbf{z})$. The ability to handle non-linear relationships sets the AE apart, allowing it to handle more intricate data sets than traditional linear techniques like PCA, thus making it a great tool for handling multimodal input data.

For the first step, we train the parameters of the AE by minimizing a combination of loss functions using backpropagation [46]. In order to maximize the quality of the disease portfolio and age interval reconstructions, the cross-entropy loss is used for the disease modality $\mathbf{x}_p$, while the mean squared error (MSE) loss is considered for the age modality $\mathbf{x}_a$:

$$\mathcal{L}_{\mathrm{A}} = \mathbb{E}\left\|\mathbf{x}_a - \mathbf{x}'_a\right\|_2^2, \tag{1}$$

$$\mathcal{L}_{\mathrm{P}} = \mathbb{E}\left[-\sum_j \left(x_{pj}\log(x'_{pj}) + (1-x_{pj})\log(1-x'_{pj})\right)\right], \tag{2}$$

where x_{pj} denotes the jth disease indicator of the disease portfolio feature vector $\mathbf{x}_p$. The combined reconstruction loss minimized is given by

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{P}} + \lambda\mathcal{L}_{\mathrm{A}}, \tag{3}$$

where the hyperparameter λ controls the weight emphasis of the age modality. Once trained, the AE encodes all observations into latent representations $\mathbf{z}$. For the second step, the k-means algorithm is run on $\mathbf{z}$ with a prespecified number of clusters k, a hyperparameter. While many state-of-the-art deep learning-based clustering methods follow an end-to-end training [25,26], the number of clusters is a dataset-dependent hyperparameter that requires careful tuning. In empirical datasets, where a "ground truth" number of clusters is typically unavailable, it is common to opt for a two-step approach [17,20], decoupling the representation learning from the clustering task which includes determining an appropriate number of clusters. This separation allows for initially learning a robust latent representation of the multi-modal input, in our situation providing valuable insights into the underlying multimorbidity mechanisms and their relationship to ageing. Employing an AE-based solution to learn a robust, unified representation of the two modalities in an unsupervised manner is thus informative, even when the input dimensionality is not large, while also offering scalability to more complex, high-dimensional datasets.

2.2.2. Limitations of autoencoder-based clustering

There are several limitations related to the AE-based feature extraction, which can hinder its usability in a clustering context. The vanilla AE lacks proper regularization, making it less robust to minor variations in the input [47]. In a clustering context, the learned representations must not be overfitted towards sporadic biases in the training data, as that would heavily influence extracted cluster patterns. A preferred approach involves a consistent, smooth and disentangled latent space. As such, we propose modifying the encoder–decoder network of the two-step approach to address these fundamental limitations.

2.3. amVAE: Age-aware multimorbidity clustering using variational AutoEncoder

We propose utilizing a variational autoencoder (VAE) [18,19] for the first step of the clustering process. VAEs have been shown capable

of learning disentangled representations of data generative factors of variation [48,49], which along with their smooth, regularized latent space make them well-suited for the clustering tasks. Additionally, VAE variants have previously been proposed for multimorbidity clustering in cross-sectional settings [13]. The VAE is architecturally similar to the vanilla AE, yet there is a fundamental difference between the two. For each input, the AE outputs a latent representation consisting of a vector, while the VAE outputs parameters of a distribution in the latent space. This is advantageous, as these parameters can be regularized towards distributions with favorable clustering properties. In order to provide a smooth structure in the latent representation, a prior distribution over the latent representation $p_\theta(\mathbf{z})$ is considered. In addition to the reconstruction loss $\mathcal{L}_{\text{rec}}$, the VAE considers a Kullback–Leibler (KL) divergence loss $\mathcal{L}_{\text{kl}}$ between the approximated posterior of the latent representation and the prior. The parameters of the VAE are obtained by minimization of the following loss:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathcal{D}_{\text{K L}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right) = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{kl}}, \quad (4)$$

where the approximate posterior is chosen as a Gaussian $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2))$ and the prior is a standard, isotropic Gaussian $p_\theta(\mathbf{z}) = N(\mathbf{0}, I)$. The tradeoff between the reconstruction and the KL-divergence losses controls the performance of the VAE. In general, the KL-divergence term acts as a regularization term, controlling the smoothness of the latent space.

When solely minimizing the MSE of the age interval modality $\mathbf{x}_a$ we do not infer strict rules between the relationship of the reconstructed variables. We designed extra auxiliary losses to provide inductive biases in order to guide the parameters of the VAE towards a more meaningful reconstruction of the age interval. For example, the VAE should never be capable of generating an age interval where the starting and/or stopping age heavily exceed or elude what is observed in the general population. Assuming the age modality is min–max normalized prior to training, we present the undershoot ($\mathcal{L}_{\text{u}}$) and overshoot ($\mathcal{L}_{\text{o}}$) auxiliary losses, designed to heavily penalize such generative behavior:

$$\mathcal{L}_{\text{u}} = \frac{1}{n}\sum_{i=1}^{n} \sum_{x_j'^{(i)} \in \mathbf{x}_a'^{(i)}} (x_j'^{(i)} \mathbb{1}_{x_j'^{(i)}<0})^2, \quad (5)$$

$$\mathcal{L}_{\text{o}} = \frac{1}{n}\sum_{i=1}^{n} \sum_{x_j'^{(i)} \in \mathbf{x}_a'^{(i)}} ((x_j'^{(i)} - 1) \mathbb{1}_{x_j'^{(i)}>1})^2, \quad (6)$$

where n determines the number of training samples.

Similarly, for the reconstructed interval to be valid, the reconstructed starting age $x_{a_1}'^{(i)}$ must be less than the reconstructed stopping age $x_{a_2}'^{(i)}$, which motivates the following valid interval ($\mathcal{L}_{\text{i}}$) auxiliary loss:

$$\mathcal{L}_{\text{i}} = \frac{1}{n}\sum_{i=1}^{n} (\max\{0, x_{a_1}'^{(i)} - x_{a_2}'^{(i)}\})^2. \quad (7)$$

We then define the total auxiliary loss as

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{u}} + \mathcal{L}_{\text{o}} + \mathcal{L}_{\text{i}}, \quad (8)$$

leading to the total loss of our proposed modified multimodal VAE:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{VAE}} + w_{\text{aux}}\mathcal{L}_{\text{aux}}, \quad (9)$$

where w_{aux} determines the weight for the auxiliary loss.

We thus present our proposed Age-aware Multimorbidity clustering using Variational AutoEncoder (amVAE) method, consisting of learning latent distributions through minimization of $\mathcal{L}_{\text{tot}}$, followed by clustering in the latent space using k-means. When clustering in the latent space, we cluster the means $\boldsymbol{\mu}_\phi$ of the approximate gaussian posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The two-step approach is illustrated in Fig. 2.

For any application of amVAE, the primary hyperparameters to tune are the weighting of the age modality relative to the disease modality, λ, and the number of clusters, k. These hyperparameters require careful tuning to the specific application. Section 3.3 describes experiments to assess consistency in the latent space, which can be used to evaluate specific choices of λ. Section 3.4 outlines experiments and metrics for optimizing both λ and k based on clinical relevance, dissimilarity and stability of the clusters.

3. Experimental setup

This section describes baselines and implementation details of models, in addition to describing the experiments and metrics used to determine an optimal clustering solution, including latent space consistency and number of clusters.

3.1. Baselines

We compare our proposed amVAE with baselines where the autoencoder architecture in the first step is replaced by alternative models, including the vanilla autoencoder (AE) and the denoising autoencoder (DAE) [47]. To make a fair comparison, the AEs and DAEs are also trained with our proposed auxiliary losses. The corruption processes considered for the DAE were masking noise for the disease portfolio modality features and gaussian noise for the age modality features.

3.2. Implementation details

Our amVAE and baselines were implemented using Pytorch [50]. All models were trained using a mini-batch size of 256 and a learning rate of 0.001 using learning rate decay. The learning rate was reduced by 50% using a reduce on plateau schedule [51], where reductions were made if validation loss did not improve in 10 epochs, repeated up to a maximum of five times. The Adam optimizer [52] was utilized. The models were trained for a total of 2000 epochs. A latent representation bottleneck of dimensionality two was utilized. Explorations with higher-dimensional latent spaces showed only slight improvements in reconstruction error. Further, two dimensions have advantages in visualizations that may further comprehension of results. An appropriate latent dimensionality is likely to be dataset-dependent and should be tuned accordingly for specific applications. Initial exploration led to an encoder–decoder setup utilizing two hidden, fully-connected layers prior to the bottleneck layer. The ReLU [53] activation function was utilized for all encoder and decoder layers except the last, which had no activation function (except for the decoder disease portfolio modality output layer, which had sigmoid activation). The age modality input data were normalized using min–max normalization according to the training data. Empirical analysis showed that $w_{\text{aux}} = 1$ was appropriate, ensuring that all of the auxiliary losses converged towards zero during training. For the DAE, initial analysis showed that a masking corruption ratio of 0.15 and gaussian noise with $\sigma = 0.05$ was appropriate.

3.3. Consistency of latent space

In order to obtain a meaningful latent space capturing the underlying structure of the input data, it is crucial for an AE-based solution to exhibit consistency across different data splits. This is to ensure that the learned representations actually extract the desired saliency of each disease that helps separating disease portfolios at different stages in life apart. Following a five fold cross validation setup, we evaluated the latent space consistency of each method, where in each split the method was trained on 80% of the data and the computed representations were based on the remaining 20%. To compare latent spaces we utilized the Bhattacharyya distance (BD), which quantifies the overlap between two distributions, providing a measure of overall similarity. The distance is given as $\text{BD} = -\log(\sum_{\mathbf{z}\in Z} \sqrt{p(\mathbf{z})q(\mathbf{z})})$ where p and q represent density estimates of two different latent spaces. The BD between each latent space was computed and the average was used to compare the different models and age modality weightings.

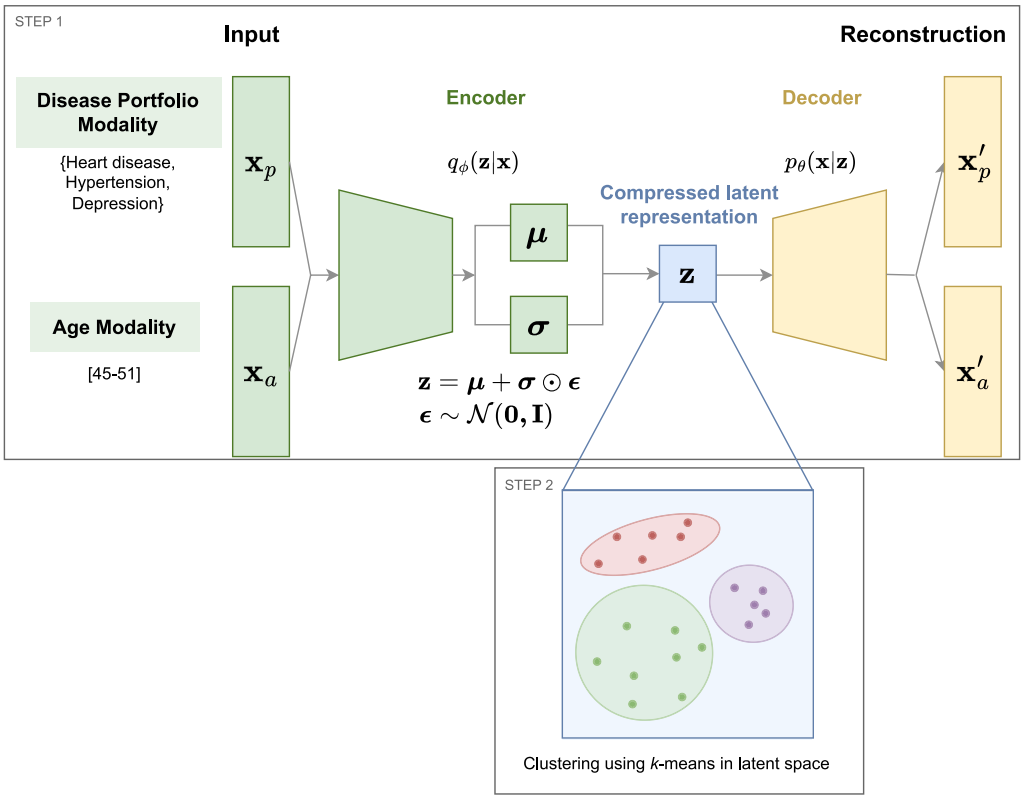


Fig. 2. Overview of our proposed two-step amVAE approach. In step one, a modified variational autoencoder is trained to compress the two modalities disease portfolio and age into a compact latent representation. In step two, clusters are extracted from the learned latent representations using k-means.

3.4. Determining the number of clusters

As the actual number of clusters k is unknown, we determine its optimal number empirically. We based our selection on three criteria. Firstly, the number of clusters should be clinically relevant, excluding a large number of clusters. Secondly, observations within a cluster should be minimally dissimilar to other observations in the same cluster, while being maximally dissimilar to observations not in the cluster. Thirdly, as data-driven determination of the number of clusters is sensitive to the choice of evaluation measure, we wanted to ensure that the optimal number of clusters was stable, meaning that going from a cluster solution of size k to a cluster solution of size $k+1$ should not provide too different of a clustering, if k is optimal.

To satisfy the first criteria, we considered cluster solutions with sizes ranging from $k = 3$ to 35, obtained from latent representations from a model trained on the full dataset. To approach a global minimum within-cluster sums of squares (WCSS) solution for each k, the Hartigan-Wong algorithm [54] was used, executing 25 random starts per run. The procedure stopped after 400 runs or if no better clustering was found within 25 runs.

To determine the optimal number of clusters with minimal dissimilarity within each cluster and maximal dissimilarity between neighboring clusters, we considered the average silhouette score [55]. The silhouette score associated with a single data point $\mathbf{z}$ is given by

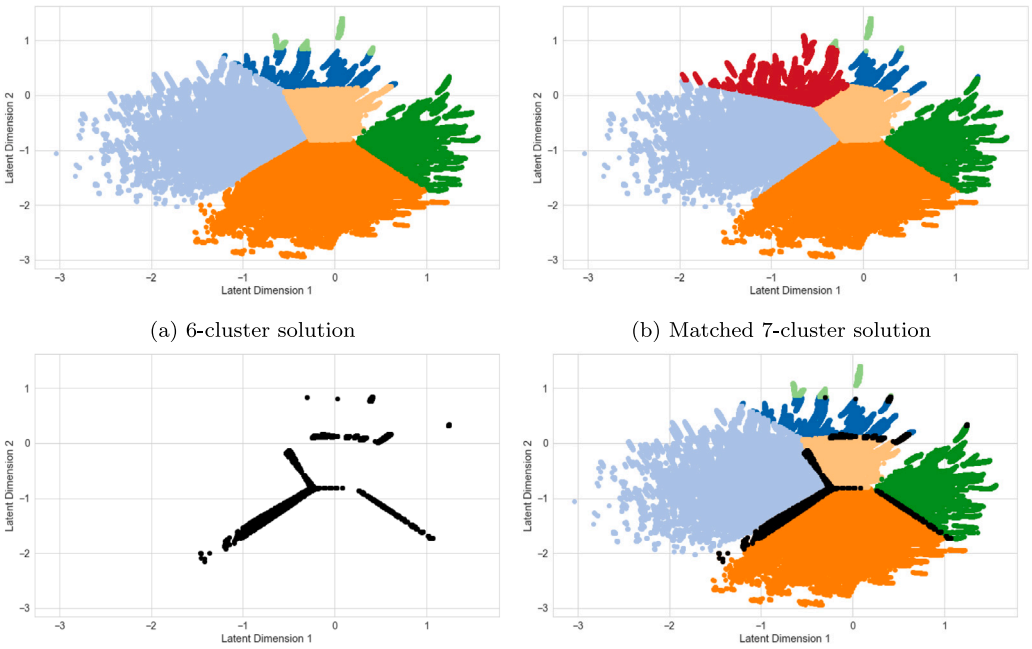
$$s(\mathbf{z}) = \frac{b(\mathbf{z}) - a(\mathbf{z})}{\max\{a(\mathbf{z}), b(\mathbf{z})\}} \tag{10}$$

where $a(\mathbf{z})$ represents the mean intra-cluster distance for $\mathbf{z}$ while $b(\mathbf{z})$ is the mean distance between $\mathbf{z}$ and the nearest cluster. By definition, the silhouette score is bounded between -1 and 1. For $s(\mathbf{z})$ to be close to 1 we require $a(\mathbf{z})$ to be much lower than $b(\mathbf{z})$. As $a(\mathbf{z})$ measures how dissimilar $\mathbf{z}$ is to its own cluster, a low value suggests the data point has a well matched cluster. Additionally, a large $b(\mathbf{z})$ implies that $\mathbf{z}$ is poorly matched to its neighboring cluster. As such, $s(\mathbf{z})$ close to 1 means that the data point is appropriately clustered.

To facilitate a stable number of clusters, we considered two metrics, comparing a size k solution to a size $k+1$ solution. The first is the Adjusted Rand Index (ARI) [56]. The general Rand Index (RI) measures the fraction of pairs of points that either both are contained in the same cluster or both are in different clusters in the two solutions. The ARI adjusts the RI for chance matches, resulting in a value between 0 and 1, where 1 indicates identical pairwise partitions, and 0 indicates independent partitions. In addition to the ARI, we consider another measure of cluster stability, inspired by the phenomenon that when going from k to $k+1$ clusters, a single cluster is usually split into two. We paired the clusters from the solution of size k to the clusters from the solution of size $k+1$ by minimizing the Euclidean distances between their corresponding cluster centroids using a modified version of the Jonker-Volgenant algorithm to solve the linear sum assignment problem [57]. As such, a single cluster in the $k+1$ solution remained unpaired, which we termed the new cluster. Any data point, where going from a size k solution, to a size $k+1$ solution led to the data point not being assigned to its matched cluster or the new cluster was termed *rim data* for k clusters, as these data points are on the rim of the clusters, allowing them to erratically change clusters when progressing in the total number of clusters [33]. We counted the frequency of rim data for each cluster solution as a second measure of stability. We considered high cluster stability attributable to low rim data frequency. An illustration of the concept is presented in Fig. 3.

4. Results

We present the quantitative results related to the experiments described in Section 3 in Section 4.1. In addition, we comprehensively analyze the best clustering solution in Section 4.2 and provide a further analysis of the importance of particular chronic conditions in a range of ablation studies in Section 4.3.



(a) 6-cluster solution (b) Matched 7-cluster solution

(c) Rim data (d) 6-cluster solution with rim data

Fig. 3. Example of the rim data concept for $k = 6$. The 6-cluster solution (a) is matched to the 7-cluster solution (b), resulting in the unpaired, new cluster colored in red. Data points in the 6-cluster solution not assigned to their matched cluster or the new cluster in the 7-cluster solution are called rim data (c, d).

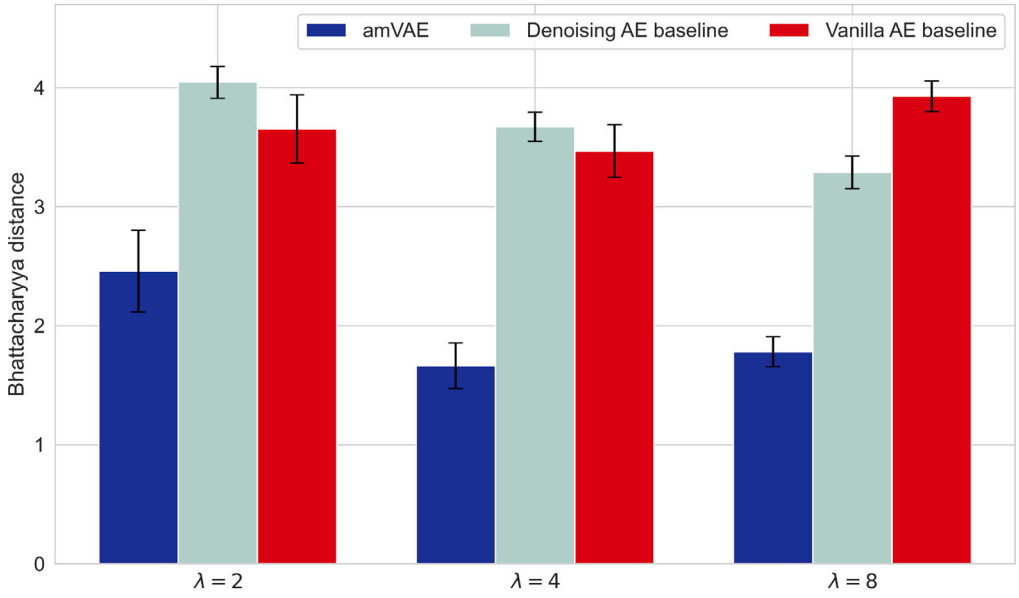


Fig. 4. Average Bhattacharyya distance (BD) between distributions of the latent representation across folds for the different models and age modality weightings (λ). Lower BD implies higher similarity between latent spaces.

4.1. Quantitative experiments

4.1.1. Comparison of latent space consistency

We compared the latent space consistency of the baseline models and our amVAE model across different age-modality weightings, as shown in Fig. 4. In general, amVAE had better latent space consistency, with its average BD considerably lower than the baselines. This consistency is likely attributed to the enforced structure and smoothness of the latent space of our amVAE model by the KL-divergence regularization. We also found that our amVAE with a moderate age-modality emphasis ($\lambda = 4$) yielded superior latent space consistency compared to both half ($\lambda = 2$) and double ($\lambda = 8$) the emphasis.

4.1.2. Clustering performance of amVAE

In Fig. 5, we present the clustering performance of our proposed amVAE model across the measures presented in Section 3.4. The model with a moderate age-modality emphasis ($\lambda = 4$) had maximal silhouette score within the range of $k = 14$ to 18, with the optimal solution found at 15 clusters. In addition, we observed that the cluster solution was most stable as indicated by both the ARI and rim data frequency within the range $k = 14$ to 16, which deviated when moving beyond the 16-cluster solution.

Regarding the model with half the age modality emphasis ($\lambda = 2$), the silhouette score was generally lower across all considered cluster

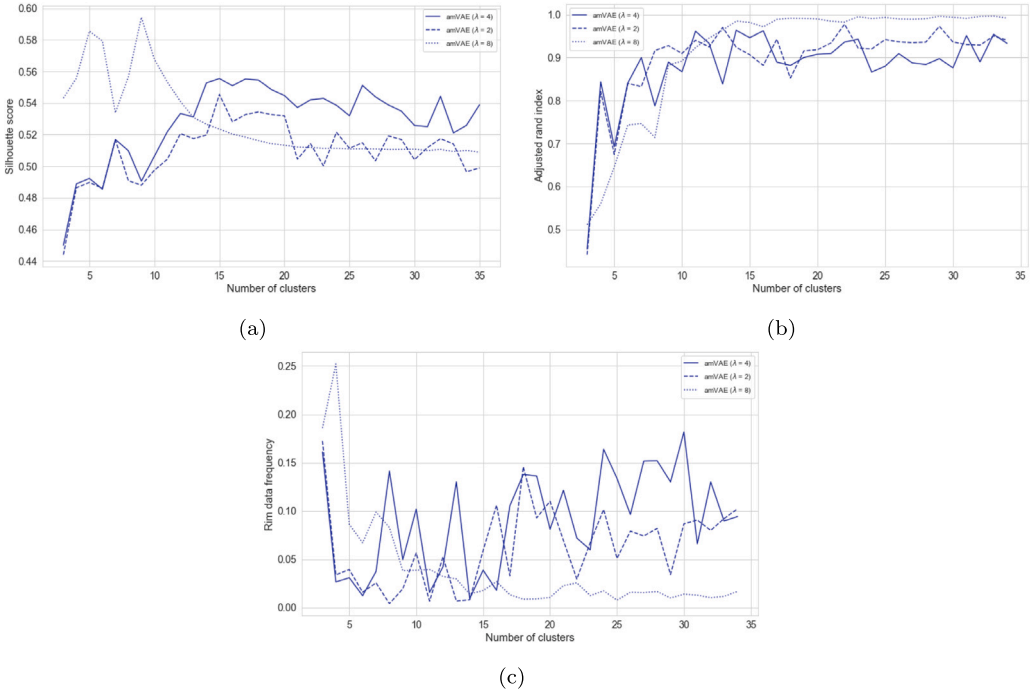


Fig. 5. amVAE clustering performance across different age weightings and number of clusters. (a) Average silhouette score. (b) Adjusted rand index comparing a k solution with a $k+1$ solution. (c) Rim data frequency.

solutions, but similarly maximized at $k = 15$. However, this model displayed worse stability around the 15-cluster solution.

Conversely, for the $\lambda = 8$ model, the silhouette score showed a distinct pattern, reaching its peak at the 9-cluster solution, gradually decreasing with an increasing number of clusters. However, the 9-cluster solution showed worse stability than the optimal $\lambda = 4$ solution. Thus, considering this worse stability along with the marginally inferior latent space consistency (Fig. 4), we did not consider this clustering solution a better choice.

We thus considered the amVAE with $\lambda = 4$ and 15 clusters the optimal solution for the chronic heart disease dataset. Apart from being optimal across our considered measures, this choice of number of clusters is clinically meaningful, as it avoids excessive complexity arising from an overly large number of clusters.

4.2. Qualitative results: Analysis and characterization of spatiotemporal clusters and the latent space

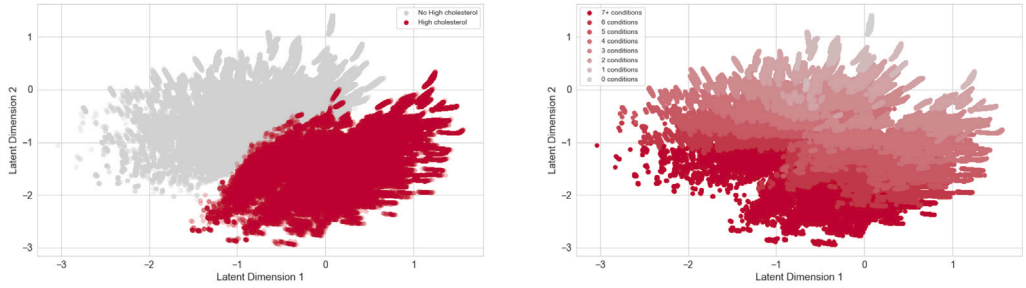
4.2.1. Structure of latent space

We carefully examined the structure of the latent representations, originating from the optimal amVAE solution. The latent space was heavily segregated by the presence of high cholesterol in the disease portfolios (see Fig. 6a), suggesting that this condition, in particular is informative for characterizing the disease trajectories. In addition, the latent space was structured according to the number of chronic conditions, where a progression from a low number of conditions to a high one was observed (Fig. 6b).

4.2.2. Characterization of clusters

Once the spatiotemporal clusters were obtained, each individual's disease trajectory was progressively embedded into the latent space and assigned clusters, allowing us to characterize the 15 clusters. The clusters were labeled based on a comprehensive analysis of three criteria: Age, level of multimorbidity, and observed/expected ratios (OEs) of chronic condition prevalence in the clusters (see Supplementary Text A.1 for a detailed labeling and analysis of the clusters). The characteristics for each of the clusters are presented in Table 1. The prevalence (or mean) of a particular variable were calculated based on individuals in the cluster. If a particular individual attained several disease portfolios within the cluster, weights for each portfolio were utilized and calculated as the ratio of time spent with the portfolio in the cluster to the total time spent in the cluster. The characteristics revealed that the patients in the different clusters were subject to varying levels of multimorbidity, ranging from an average number of conditions of 0 to 6.8. In general, the older clusters were subject to higher levels of multimorbidity, yet the clustering resulted in an early multimorbid high cholesterol infused cluster (the "Early multimorbid high cholesterol and heart disease" cluster). General chronic condition prevalence among the clusters revealed that the disease trajectories of the HD patients were heavily prone to also include hypertension and high cholesterol. The mental health, musculoskeletal and respiratory conditions were also largely prevalent, especially in the later clusters (see Supplemental Figure B.10 for a visualization). The clusters also exhibited variation in the distribution of sex and educational attainment levels (Supplemental Figures B.11–B.12).

Fig. 7 shows a heatmap of OEs for the prevalence of each condition in the clusters. The OE was calculated as the prevalence of the chronic condition in the cluster divided by the general prevalence of the condition in the population in the mean age interval of the cluster. The OEs revealed essential differences among the disease burden of the different clusters. For example, the main difference between the two clusters "Late severe complex multimorbidity" and "Late complex multimorbid (diabetes and high cholesterol underrepresented)", was the underrepresentation of high cholesterol and diabetes in the latter cluster. The OEs similarly revealed essential insights into the role of the different chronic conditions among the clusters. Conditions such as high cholesterol, osteoporosis and dementia were generally either overrepresented or underrepresented in the clusters, showing them central to cluster formation. In contrast, hypertension, although highly prevalent in the HD population, was typically prevalent at its expected level. The diabetes condition was also informative, as it was typically present in the cluster along with a few other conditions (i.e. heart disease or hypertension in the cluster "Late multimorbid heart disease, diabetes and hypertension") or underrepresented when many other conditions were overrepresented (Fig. 7).



(a) (b)

Fig. 6. Latent space from optimal amVAE model ($\lambda = 4$) colored by prevalence of high cholesterol (a) and number of conditions in disease portfolio (b).

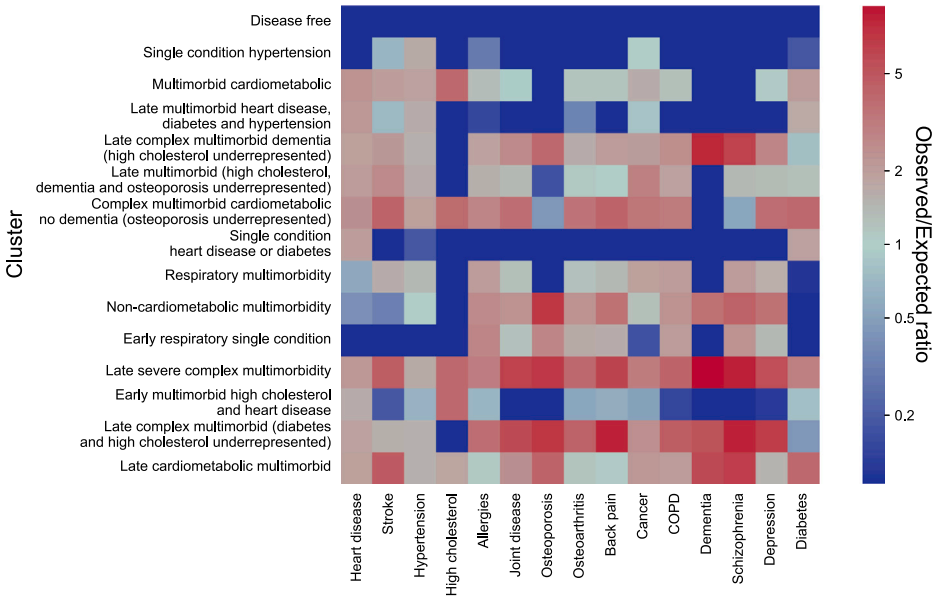


Fig. 7. Observed/Expected ratios for the prevalence of each condition in each cluster.

4.2.3. Cluster transitions

To obtain further insight into the spatiotemporal structure of the disease trajectories, we analyzed cluster transitions within the chronic heart disease population. Table 2 displays the 10 most common cluster trajectories. The HD population's cluster trajectories were heterogeneous, as the ten most frequent trajectories only accounted for 22.2% of the population. The cluster "Multimorbid cardiometabolic" emerges as a central multimorbidity cluster, displaying that the combination of heart disease, hypertension and diabetes is a common stage in the progression of multimorbidity (mean age interval 65.7–71). In addition, the single condition clusters "Single condition hypertension" and "Single condition heart disease or diabetes" commonly follows the "Disease free" cluster.

Fig. 8 visualizes the age distribution of the clusters, along with the most common transitions between them. The figure reveals a segregation of common pathways in the population. Broadly, three main trajectory highways are illustrated, an upper, middle and lower pathway. The upper pathway was characterized by a development of respiratory and non-cardiometabolic conditions prior to a late complex multimorbidity state. The middle HD pathway was centered around hypertension and diabetes, with the lower pathway harbored by high cholesterol. Notably, the lower pathway clusters had longer mean intervals, indicating that individuals typically stay in the clusters for a longer period of time before transitioning to a new cluster. The clusters "Single condition heart disease or diabetes" and "Late multimorbid heart disease, diabetes and hypertension" manifested themselves as clusters where there is a moderate likelihood of transitioning into the high cholesterol infused clusters, from the non-cholesterol clusters. In a subsequent analysis, we constructed similar cluster trajectory visualizations for specific subpopulations of the HD population. These were the ischaemic heart disease, heart failure and cardiac therapy drug consuming patients (see Supplemental Figure B.13). For the subpopulations it was observed that the ischaemic heart disease patients predominately followed the lower, high cholesterol harbored pathway, while the heart failure patients predominantly appeared in the upper and middle pathways.

4.3. Ablation studies

In order to understand the contribution of each chronic condition on the structure of the latent space, we conducted an ablation experiment where we removed a single condition from the dataset before training of the VAE. We evaluate the distortion of the resulting latent space by calculating its similarity to the original latent space using the BD. The results are shown in Table 3. The results show that the absence of high cholesterol leads to a very dissimilar latent space, whereas the latent space obtained when e.g. cancer is removed is much more similar to the original latent space, with much of the structure remaining (Fig. 9 provides a visualization). In addition to high cholesterol, the ablation of the conditions COPD and hypertension resulted in a latent space with higher dissimilarity, whereas the ablation of schizophrenia and dementia showed slight dissimilarity to the original latent space.

Table 1
Characteristics of the 15 clusters. Presented values are % (count) or mean (SD).

Cluster	Disease free	Single condition hypertension	Multimorbid cardiometabolic	Late multimorbid heart disease, diabetes and hypertension	Late complex multimorbid dementia (high cholesterol underrepresented)	Late multimorbid (high cholesterol, dementia and osteoporosis) underrepresented)	Complex multimorbid cardiometabolic no dementia (osteoporosis underrepresented)	Single condition heart disease or diabetes
Chronic Conditions								
Heart disease	0[a] (0)	4.3 (14850)	81.6 (203851)	92.1 (285255)	91.4 (188958)	90.9 (229170)	91.9 (148951)	80 (175972)
Stroke	0[a] (0)	4.5 (15433)	11.5 (31081)	5.5 (17267)	20.8 (49117)	21.5 (56966)	29.7 (49711)	0 (102)
Hypertension	0[a] (0)	84.1 (290077)	84.4 (199417)	86.5 (265534)	89.2 (177418)	93.2 (233544)	93.4[b] (145171)	9.4 (24686)
High cholesterol	0[a] (0)	0[a] (0)	100[b] (233837)	0.8 (2388)	0.7 (1612)	0.1 (354)	100[b] (154118)	0[a] (0)
Allergies	0[a] (0)	5.3 (18228)	19.9 (51360)	2.6 (8599)	34.5 (70471)	28 (71468)	43.6 (70383)	0 (28)
Joint disease	0[a] (0)	0 (136)	1.2 (3514)	0.1 (353)	4.5 (9522)	2.3 (5768)	5.3 (9519)	0[a] (0)
Osteoporosis	0[a] (0)	0[a] (0)	0[a] (0)	0.2 (698)	41.9 (86072)	1.4 (3501)	2.7 (4240)	0[a] (0)
Osteoarthritis	0[a] (0)	0.2 (672)	7.2 (20520)	2.6 (8083)	14.6 (30115)	9 (22856)	23.7 (40085)	– –
Back pain	0[a] (0)	0.2 (810)	5.5 (15125)	0.5 (1503)	11.1 (22752)	5.3 (13297)	20 (34032)	0[a] (0)
Cancer	0[a] (0)	5.3 (18252)	7.5 (25319)	5.1 (16360)	14.4 (34359)	19.1 (52944)	18.7 (35053)	0.3 (870)
COPD	0[a] (0)	0.3 (952)	10.2 (28702)	1.0 (3256)	26.8 (57249)	20.1 (51188)	30 (51494)	0[a] (0)
Dementia	0[a] (0)	0[a] (0)	0[a] (0)	0[a] (0)	18.4[b] (41237)	0.1 (176)	0[a] (0)	0[a] (0)
Schizophrenia	0[a] (0)	0[a] (0)	0[a] (0)	0.1 (154)	7.4 (17577)	1.4 (3616)	0.5 (838)	0[a] (0)
Depression	0[a] (0)	0.3 (994)	8.2 (21261)	0.3 (855)	26.3 (53977)	11.8 (29613)	30.5 (50705)	0[a] (0)
Diabetes	-[a] –	2.0 (6942)	20.1 (52975)	18.7 (63425)	8.5 (19401)	13.8 (37894)	41.1 (67519)	21.8 (48292)
Multimorbidity								
2+ conditions	0[a] (0)	6.5 (22612)	100[b] (233837)	100[b] (306122)	100[b] (196458)	100[b] (249774)	100[b] (154118)	11.5 (32528)
Complex multimorbidity*	0[a] (0)	0 (16)	6.1 (17130)	0.1 (341)	64.0 (138285)	9.1 (29092)	53.2 (94002)	0[a] (0)
Number of Conditions	0[a] (0)	1.1 (0.2)	3.6 (0.5)	2.2 (0.4)	4.1 (0.8)	3.2 (0.4)	5.3 (0.6)	1.1 (0.3)
Age								
Starting age	61.8 (15.6)	66.5 (13)	65.7 (10.9)	71.5 (12.6)	77.6[b] (11.4)	73.4 (12)	69.6 (10.6)	66.1 (14.5)
Stopping age	65.8 (14)	69.6 (12.8)	71 (10.8)	74.4 (12.6)	80.3[b] (11.3)	75.9 (11.9)	73.8 (10.2)	68.3 (13.9)
Sex								
Female	46.9 (359173)	50.5 (174150)	36 (84085)	45.6 (139742)	62.6 (122954)	48.6 (121327)	41.2 (63541)	38.8 (84439)
Education								
None (≤ 10 years)	33.6 (257468)	34.8 (119870)	38.4 (89801)	31.6[a] (96852)	34.1 (67020)	34.7 (86785)	42 (64798)	31.7 (68962)
Short (11–14 years)	29.8 (228050)	29.3 (101028)	42.2 (98632)	24.7 (75663)	22.2[a] (43699)	25.8 (64391)	40.1 (61759)	29.7 (64615)
Medium (15–16 years)	5.6 (42728)	5.8 (20078)	8.6 (20095)	4.8 (14765)	4.3[a] (8413)	5.0 (12519)	7.6 (11699)	6 (12999)
Long (≥ 17 years)	4.1 (31039)	4.2 (14559)	5.7 (13406)	3.5 (10635)	3.1[a] (6115)	3.7 (9299)	4.9 (7619)	4.2 (9065)
Missing	26.9 (206006)	25.9 (89183)	5.1 (11903)	35.3[b] (108207)	36.2 (71211)	30.7 (76780)	5.3 (8243)	28.4 (61784)

Cluster	Respiratory multimorbidity	Non-cardiometabolic multimorbidity	Early respiratory single condition	Late severe complex multimorbidity	Early multimorbid high cholesterol and heart disease	Late complex multimorbid (diabetes and high cholesterol underrepresented)	Late cardiometabolic multimorbid
Chronic Conditions							
Heart disease	23.4 (49890)	16.4 (25864)	0[a] (0)	94.5[b] (85691)	54.3 (63559)	82.1 (59794)	86.9 (73779)
Stroke	11.3 (24252)	2.3 (3684)	– –	39.3 (36913)	0.9 (1051)	13.5 (11351)	41.1[b] (35604)
Hypertension	70.2 (149946)	50.7 (69628)	0[a] (0)	94 (81416)	27.5 (32448)	84.5 (54423)	87.4 (72614)
High cholesterol	0[a] (0)	0[a] (0)	0[a] (0)	98.1 (84140)	100[b] (108303)	0 (16)	47.7 (39234)
Allergies	35.2 (75618)	42.6 (54031)	43.3 (71179)	52.3 (46655)	10.7 (13508)	64.1[b] (41293)	19.4 (16717)
Joint disease	1.9 (3953)	3.4 (4463)	1.4 (2290)	10.2[b] (9688)	0.1 (92)	9.2 (6332)	3.9 (3411)
Osteoporosis	0.5 (1060)	42.2 (53014)	9.9 (16235)	57.7 (52080)	0[a] (0)	62.8[b] (41463)	36.8 (30558)
Osteoarthritis	8.4 (18184)	16.5 (21757)	8.8 (14494)	31.8 (29301)	2.7 (4186)	36.1[b] (23887)	9.7 (8402)
Back pain	7.5 (15964)	18 (23116)	8.7 (14219)	31.3 (29050)	3.1 (4421)	42.3[b] (27713)	5.7 (4936)
Cancer	10.6 (24317)	7.1 (10366)	0.6 (1592)	20.4[b] (21446)	1.8 (2966)	16.3 (12786)	15.1 (14539)
COPD	19.5 (41928)	22.4 (30251)	14.2 (23992)	43 (41017)	1.0 (1434)	48.4[b] (33129)	21.9 (19574)
Dementia	0[a] (0)	2.5 (3194)	0[a] (0)	14.4 (15006)	0[a] (0)	8.5 (7688)	8.7 (7341)
Schizophrenia	2.2 (4678)	4.4 (5971)	2.7 (4513)	8.8[b] (8754)	0[a] (0)	8.7 (6576)	6.9 (5919)
Depression	13.7 (29138)	29.5 (37861)	11.8 (19370)	49.2 (44260)	1.1 (1388)	58.5[b] (38194)	13.7 (11965)
Diabetes	1.3 (2828)	0.4 (677)	0.1 (341)	31.2 (29025)	7.9 (10324)	4.6 (3949)	43.7[b] (36502)
Multimorbidity							
2+ conditions	100[b] (212569)	98.1 (120749)	1.7 (4696)	100[b] (85698)	92 (106994)	100[b] (62607)	100[b] (82147)
Complex multimorbidity*	3.7 (11477)	41.8 (67612)	0[a] (0)	90.4 (79364)	0.1 (220)	96.8[b] (61353)	37.2 (34530)
Number of Conditions	2.1 (0.2)	2.6 (0.6)	1.0 (0.1)	6.8[b] (1.1)	2.1 (0.5)	5.4 (1)	4.5 (0.7)
Age							
Starting age	68 (12.9)	69.4 (13.8)	60.8 (14.8)	74.6 (10)	60.8[a] (11)	75.2 (12.2)	74.7 (11)
Stopping age	70.7 (12.6)	72.8 (13.1)	64.9[a] (13.8)	78.2 (9.7)	65.5 (11)	78.7 (12)	77.4 (10.7)
Sex							
Female	51.7 (109991)	65.3 (79969)	51.5 (84295)	59.1 (50608)	30.7[a] (33290)	70.6[b] (44172)	52.7 (43279)
Education							
None (≤ 10 years)	36.8 (78269)	36.6 (44753)	35.1 (57409)	47[b] (40275)	34.2 (37086)	40.1 (25130)	41.4 (34049)
Short (11–14 years)	31.6 (67108)	30.5 (37308)	36.4 (59536)	35.2 (30175)	45.4[b] (49216)	27.5 (17236)	28.9 (23735)
Medium (15–16 years)	6.3 (13319)	6.1 (7518)	7.5 (12213)	6.5 (5587)	9.7[b] (10490)	5.4 (3398)	5.2 (4276)
Long (≥ 17 years)	4.5 (9560)	4.1 (5031)	5.4 (8846)	4.2 (3575)	6.6[b] (7164)	3.5 (2221)	3.7 (3014)
Missing	20.8 (44313)	22.7 (27779)	15.6 (25526)	7.1 (6086)	4.0[a] (4347)	23.4 (14622)	20.8 (17073)

* At least three conditions affecting at least three body systems.

[a] Lowest across clusters.

[b] Highest across clusters.

Table 2
Most common full cluster trajectories.
MM = Multimorbid, UR = Underrepresented, HD = Heart disease, HC = High cholesterol, HT = Hypertension, OP = Osteoporosis, DEM = Dementia, DIA = Diabetes.

Rank	Cluster trajectory					%	(count)
	First	Second	Third	Fourth	Fifth		
1	Disease free	Single condition hypertension	Late MM HD, DIA and HT	Death		4.06	(31122)
2	Disease free	Single condition HD or DIA	Late MM HD, DIA and HT	Death		2.94	(22550)
3	Disease free	Single condition hypertension	Late MM HD, DIA and HT	Late MM (HC, DEM and OP UR)	Death	2.76	(21158)
4	Disease free	Single condition hypertension	Respiratory multimorbidity	Late MM (HC, DEM and OP UR)	Death	2.12	(16265)
5	Disease free	Single condition hypertension	Late MM HD, DIA and HT	Multimorbid cardiometabolic		2.05	(15714)
6	Disease free	Single condition HD or DIA	Late MM HD, DIA and HT	Late MM (HC, DEM and OP UR)	Death	2.00	(15336)
7	Disease free	Single condition HD or DIA	Death			1.98	(15204)
8	Disease free	Single condition HD or DIA	Early MM HC and DIA			1.75	(13415)
9	Disease free	Single condition HD or DIA	Early MM HC and DIA	Multimorbid cardiometabolic		1.33	(10166)
10	Disease free	Single condition hypertension	Early MM HC and DIA	Multimorbid cardiometabolic		1.21	(9276)

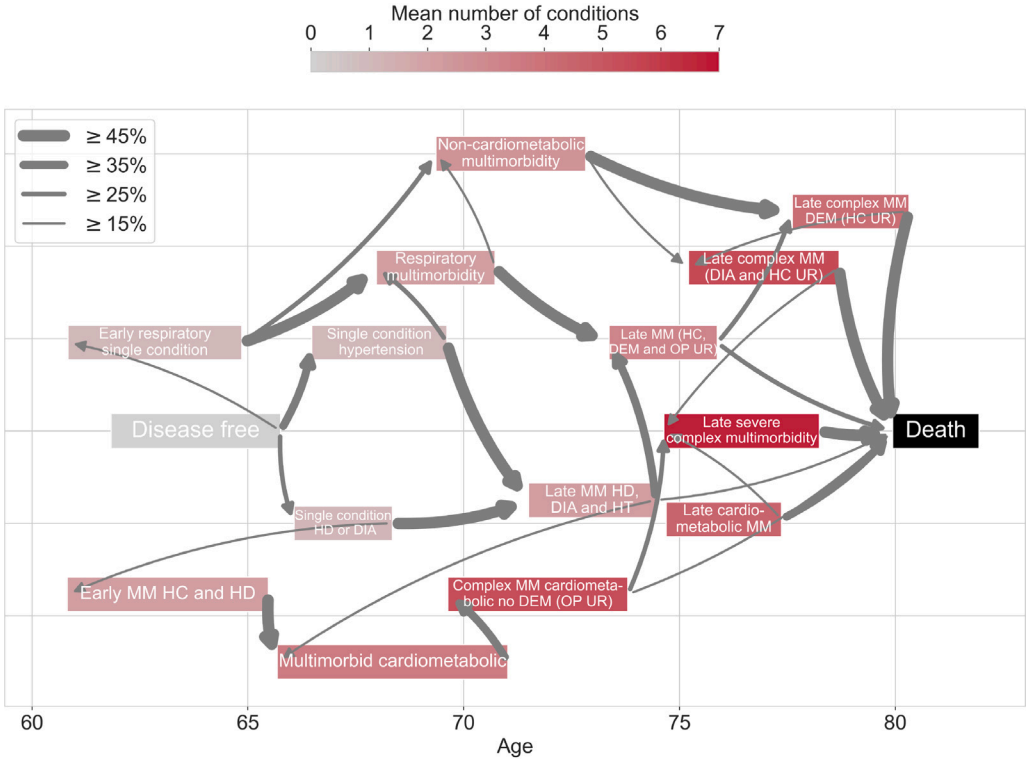
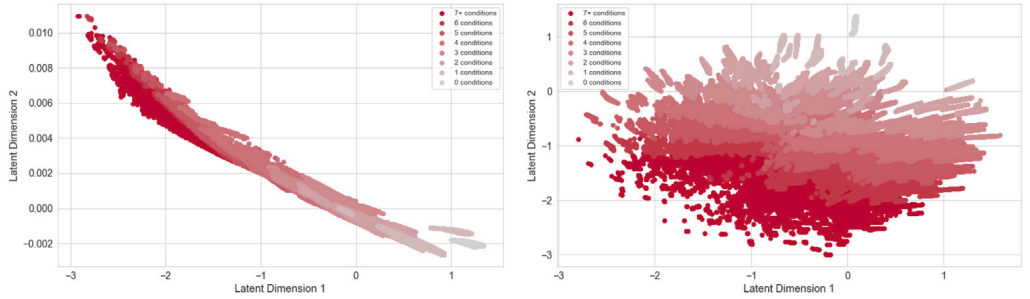


Fig. 8. Cluster trajectory highways. The figure illustrates mean age intervals for each cluster, corresponding to the mean starting age and stopping age of each individual who enters or leaves the cluster. Death is included as a separate state at the mean mortality age. The arrows pointing out of a cluster indicate the proportion of individuals in the cluster transitioning to a specific new cluster. The width of the arrow corresponds to the magnitude of the proportion. The color of a cluster corresponds to the mean number of conditions in the specific cluster (red more severe).
MM = Multimorbid, UR = Underrepresented, HD = Heart disease, HC = High cholesterol, HT = Hypertension, OP = Osteoporosis, DEM = Dementia, DIA = Diabetes.



(a) (b)

Fig. 9. Latent space obtained from training on dataset without (a) high cholesterol (b) cancer.

Table 3
Results of ablation study experiment. The similarity between the latent space resulting from having removed a particular condition from the training dataset and the full dataset (VAE $\lambda = 4$).

	Bhattacharyya Distance
Stroke	1.58
Hypertension	2.65
High cholesterol	5.04
Allergies	2.22
Joint disease	1.57
Osteoporosis	1.95
Osteoarthritis	1.63
Back pain	2.30
Cancer	1.48
COPD	2.65
Dementia	1.48
Schizophrenia	1.46
Depression	2.44
Diabetes	1.66

5. Discussion

In this paper we elucidated how multimorbidity clusters appear in different stages of life in the chronic heart disease population. We did this in an unsupervised fashion through amVAE, our proposed two-stage VAE-based approach for temporal disease-based clustering. In the first stage, a modified VAE was trained on the two modalities, the disease portfolio and its associated age interval, embedding them into a joint latent representation. The VAE proved to be vital for obtaining a consistent, regularized latent space well suited for clustering. In the second stage, clusters were formed in the latent space by the k-means algorithm, resulting in a spatio-temporal clustering, where disease portfolios occurring at similar stages in life were clustered together. Our proposed amVAE framework addresses the gap of embedding age as a continuous concept in the temporal disease-based clustering. Additionally, to the best of our knowledge, we performed temporal disease-based clustering for the first time on a nationwide population with important information from both the primary and secondary sector of the healthcare system. The framework was validated on a nationwide longitudinal dataset of chronic heart disease patients, analyzing their whole disease trajectory based on information of chronic conditions arising from diagnostic algorithms. In this population, the extracted latent space was segregated mainly by the presence of high cholesterol, showing it essential for differentiating between various types of multimorbidity in chronic heart disease patients. This was visible in the 15 extracted clusters, which were either over- or under-represented by high cholesterol, along with osteoporosis and dementia. The 15 clusters had different disease and age profiles, showcasing variations in disease burden across clusters in both type and number of conditions. In addition, our methodology discovered a high cholesterol infused multimorbidity stage, arising early in the disease trajectories (the "Early multimorbid high cholesterol and heart disease" cluster). Individuals within this cluster were likely to transition into some of the most severe clusters with a mixture of serious conditions, as opposed to other clusters with similar age profiles, but lower burden of disease. Hence, patients sharing characteristics akin to those in this cluster (high cholesterol and heart disease in the age range 60–65) represent prime candidates for preventive treatments. For the overall structure of the chronic heart disease trajectories, our analysis revealed three primary cluster pathways; one with respiratory complications, another revolving around hypertension and diabetes, and a third harbored by high cholesterol. It is worth noting the imbalanced sex composition within certain clusters. For example, female predominance is visible in the clusters labeled "Late complex multimorbid dementia (high cholesterol underrepresented)" and "Non-cardiometabolic multimorbidity". This predominance is even more evident in the "Late complex multimorbid (diabetes and high cholesterol underrepresented)" cluster. These clusters share common characteristics, such as an overrepresentation of mental health conditions and osteoporosis, along with an underrepresentation of high cholesterol. These findings underscore the presence of sex-specific differences in the life course trajectories of HD patients [43].

5.1. Transitions between clusters

In our application of amVAE on HD patients, we revealed three primary cluster transition pathways (Fig. 8). The upper pathway was notably complicated by respiratory conditions, often occurring before or alongside HD. This pathway was the primary pathway for heart failure patients, which is consistent with the increased risk of heart failure associated with tobacco smoking [58]. Given the overrepresentation of COPD in this pathway, we hypothesize that many of these patients are smokers, whose smoking habits contribute to the development of heart failure. The lower pathway consisted of high cholesterol over-represented clusters and was primarily occupied by ischaemic heart disease patients. As high cholesterol is an important risk factor for ischaemic heart disease, also this finding falls well in line with existing research [59]. High cholesterol also played a crucial role in the latent representations of our amVAE model, as the condition segregated the latent space, separating observations with high cholesterol from those without (Fig. 6). This emphasizes high cholesterol as highly informative for subtyping the different HD multimorbidity profiles at different stages in life. While high cholesterol is considered an important risk factor for heart disease [60], it is interesting that it manifests itself in an early multimorbidity cluster in the HD population (the "Early multimorbid high cholesterol and heart disease" cluster, Table 1, Fig. 8). In this cluster, heart disease only has a prevalence of 54.3%, yet high cholesterol on average is present with one other chronic condition. Following the most common cluster transitions (Fig. 8), we observed that these individuals are likely to transition into some of the most complex multimorbidity clusters of highest average number of conditions at a later stage. Despite this, the duration of which these individuals remain in single clusters are more prolonged, compared to other clusters at similar stages in life. As such, the average development of conditions is slower for these individuals, and one might hypothesize that they represent a less frail part of the population who are more aware of their conditions, being diagnosed and treated at an earlier stage. The cluster "Early multimorbid high cholesterol and heart disease", despite being the sole early multimorbidity cluster also had the highest proportion of individuals with short, medium and long education across the clusters (Table 1). As higher educational attainment is associated with a healthier lifestyle [61], this might support our interpretation of these individuals being more aware of their health. On the other hand, this group of patients may receive effective medical treatment, resulting in the slower progression towards more severe clusters. Our results showcased the clusters "Single condition heart disease and diabetes" and "Late multimorbid heart disease, diabetes and hypertension" as facilitating transitions into the high cholesterol infused part of the cluster trajectories (Fig. 8). These clusters generally consisted of diabetes, heart disease and hypertension, conditions which all share common risk factors for high cholesterol such as obesity, smoking and physical inactivity [62] and can cause inflammation, potentially leading to an accumulation of cholesterol on the artery walls [63].

5.2. Advancements from previous temporal disease-based approaches

Our proposed amVAE address the issue of embedding continuous age-time into the clustering, which existing temporal disease-based clustering approaches were unable to. In the study conducted by Vetrano et al. [21] consisting of multiple separate cross-sectional clustering analyses, the timing of diagnoses is not preserved, as they can

occur at any arbitrary point in time between the separate analyses. The lack of knowledge about the timing of diagnoses introduces uncertainty regarding the order of conditions and the progression of clusters over time. It is possible that an intermediate, unobserved cluster could have existed between the data-collecting points in time, which further complicates the interpretation of cluster progression during the study period. However, even the factorization-based methods [22,23] that leverage EHRs do not explicitly model age as a temporal concept. This omission is essential in multimorbidity research, given the well-established association between age and multimorbidity [42,64]. Age plays a crucial role in multimorbidity dynamics, as many underlying diagnoses occur at different stages of an individual's life, leading to heterogeneous prevalences of conditions across different age groups [65]. Additionally, it is essential to note that the constructed temporal dimension of the factorization methods assumes a uniform time interval, typically yearly, between disease observations. This simplification may not adequately account for possible variations in the rate of disease progression, as important changes and transitions occurring over short age intervals might be overlooked, limiting a comprehensive understanding of the dynamics of multimorbidity. Additionally, existing studies, such as [21,22], both rely on data from a few thousand individuals. In contrast, while [23] utilize a large population covering millions of patients, they solely consider ICD-10 diagnosis codes, overlooking the potentially valuable information about chronic conditions which can be derived from the patient's prescribed medication. These limited populations restrict the generalizability and potential insights which can be gained from their discovered temporal disease clusters. In this study, we combated this by utilizing diagnostic algorithms targeting the entire population, thus moving beyond solely considering the secondary healthcare sector.

5.3. Implications

Characterizing and analyzing temporal disease-based clusters of multimorbidity in subpopulations such as the HD population is essential for understanding the epidemiology of multimorbidity and the patterns for development of concomitant more conditions in order to create personalized patient trajectories in the healthcare system. The insights we have acquired from mapping the distinct phases in HD trajectories can inform the design of more efficient treatment and prevention strategies. Further, this information might inform development of clinical guidelines for HD and the most common co-occurrent conditions. At the population level, preventive and general healthcare strategies should be guided by an understanding of the different cluster pathways. Likewise, at the clinical level, the provided knowledge suggests the need for personalized treatment approaches tailored to patients within different clusters. As most cardiovascular patients are multimorbid [43], clinicians might improve their insight into the nature of development of concomitant conditions in the HD patient population by consulting the clusters. In addition, understanding which clusters a person is likely to transition to, could guide the clinician in focusing on which possible chronic conditions they should be aware of, thus focusing on specific preventive treatments. Lastly, clinical guidelines taking into account multimorbidity [66] for HD could be updated with information on the individual clusters, as recommended treatments are likely to differ among the different stages in the HD trajectory. In summary, the cluster characteristics provided in this study may be informative for population-based care of individuals suffering from HD, development of clinical guidelines in HD and hypothesis generation on important disease portfolio progression mechanisms.

5.4. Cluster characteristics

In our presentation of characteristics related to the various clusters (Table 1), we computed prevalences (or means) on an individual-centered basis, where each patient contributed equally to the total weight for each prevalence. This approach allows us to interpret the characteristics (e.g. chronic conditions) as follows: The prevalence represents the probability that a randomly selected patient has a specific condition while they belong to the cluster. Alternatively, we could have employed a time-based prevalence, computed as the time (in years) a characteristic is present in the cluster divided by the total time spent in the cluster. In this case, the prevalence would correspond to the expected frequency of the condition at any random point in time within the cluster. In either case, interpretations may have some ambiguity, as they attempt to summarize something that can dynamically change over time with a single numeric value. However, we opted for an individual-centered reporting as this approach offers a more intuitive interpretation regarding the organization of different patient pathways in the healthcare system. Additionally, time-based prevalences did not lead to significant changes in the figures and cluster labels.

5.5. Limitations and future work

Our proposed work does not come without limitations. While, our proposed amVAE is purely unsupervised, it comes with a number of hyperparameters that need to be tuned. It is essential to tune the λ parameter in order to allow the VAE to properly reconstruct both modalities, without collapsing solely into either. For our dataset, the λ which exhibited the best latent space consistency also balanced the two losses, however, this might not be the case for other datasets. It is therefore important to conduct relevant experiments in order to ensure consistency and balanced modalities (see Fig. 4). In addition, determining the optimal number of clusters is important. We argue that for this task it is essential to consider stability measures in conjunction with traditional clustering indices. While we have worked with chronic conditions based on algorithmic diagnoses, we acknowledge that many of the considered conditions such as heart disease and cancer come in heterogeneous forms. As an example, our considered HD diagnosis includes both ischaemic and heart failure patients. However, despite the amVAE model not being aware of the different forms of HD during training, the various HD forms are recovered and expressed in distinct parts of the cluster trajectories (see Supplemental Figure B.13). This is interesting because it indicates that the various patient groups experience distinct health trajectories, influenced by their comorbidities and the age intervals associated with the different disease portfolios. For future work, we intend to work directly with diagnoses at an ICD-10 level in conjunction with ATC medicine codes. Using these two data sources as extra modalities, the clustering could be based on a more nuanced view of each patient's disease trajectory. In terms of model architecture, there are several possibilities which potentially could enhance the performance. One possibility is to model the prior distribution of the latent space as a Gaussian mixture, as suggested in [67]. Such an approach could promote further disentanglement of the underlying patterns in the data. Nevertheless, it would introduce the challenge of determining the number of components in the prior, which becomes difficult when the true number of clusters remains unknown. Alternatively, we could explore a temporary clustering of low order on the latent representations of the VAE during training, analogous to the concept proposed in [68]. By calculating e.g., a clustering silhouette score and incorporating it into an additional loss term, the VAE could be encouraged to extract representations which are distinctly separated into meaningful groups in the latent space. We defer this to later research. Other lines of future research could involve incorporating the cross-attention mechanism, as seen in transformer models [69], between the intermediate embeddings of the disease and age modalities within the encoder network. This could lead to augmented VAE representations, potentially with improved ability to capture complex interactions between the two modalities. While we demonstrated our methodology on a population of chronic heart disease diagnosed individuals in order to obtain insights on their associated multimorbidity progression, further studies could apply the methodology on either (a) a subpopulation conditioned on a different multimorbidity-prone chronic condition such as diabetes [70] or COPD [71] or (b) the general population.

6. Conclusion

In this study, we introduce amVAE, a novel two-step approach for temporal disease-based clustering of multimorbidity patterns. Clusters extracted by our methodology can serve as catalysts for generation of new hypotheses on the epidemiology of multimorbidity, how chronic conditions cluster over time, and patient transitions between clusters. These patterns can serve as evidence to support the much needed development of clinical guidelines on multimorbidity. Through a comprehensive analysis of nationwide disease trajectories encompassing all Danish chronic heart disease patients from 1995 to 2015, our amVAE model revealed an underlying multimorbidity structure segregated by a few key conditions. In particular, the presence of high cholesterol, osteoporosis and dementia were pivotal in characterizing the separate clusters within the chronic heart disease population. Our analysis of transitions between clusters revealed three main trajectory pathways among chronic heart disease patients: a pathway characterized by respiratory multimorbidity common in heart failure patients, another complicated mainly by hypertension and diabetes, with a third characterized by early multimorbidity and dominated by high cholesterol, typical for ischaemic heart disease patients. Hence, our findings underscore the imperative for a paradigm shift in healthcare for patients with chronic heart conditions, emphasizing integrated, person-centered care over isolated silo-based approaches.

CRediT authorship contribution statement

Nikolaj Normann Holm: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thao Minh Le:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Anne Frølich:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Ove Andersen:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation. **Helle Gybel Juul-Larsen:** Writing – review & editing, Validation, Resources, Investigation, Data curation. **Anders Stockmarr:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Svetha Venkatesh:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Ethical approval and consent to participate

The Danish national registries are protected by the Danish Data Protection Act and can only be accessed after application and subsequent approval. This study did not require additional approval from the Danish Research Ethics Committees, as it solely involved the use of national registry data, which is exempt under the Scientific Ethical Committees Act. No informed consent was required.

Declaration of competing interest

The authors report no potential conflict of interests.

Acknowledgments

We would like to acknowledge the Applied Artificial Intelligence Institute at Deakin University for hosting Nikolaj Normann Holm on his external research stay, under which this research was performed.

This research was performed as part of the Clinical Academic Group Prognostication of Acute Recovery Capacity - in an Aging Population (ACUTE-CAG) funded by the Greater Copenhagen Health Science Partners (GCHSP), Denmark. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109632.

References

[1] J.M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, M. Roland, Defining comorbidity: Implications for understanding health and health services, Ann. Fam. Med. 7 (4) (2009) 357–363, http://dx.doi.org/10.1370/afm.983, arXiv:https://www.annfammed.org/content/7/4/357.full.pdf, URL https://www.annfammed.org/content/7/4/357.

[2] S.H. van Oostrom, R. Gijsen, I. Stirbu, J.C. Korevaar, F.G. Schellevis, H.S.J. Picavet, N. Hoeymans, Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: data from general practices and health surveys, PLoS One 11 (8) (2016) e0160264.

[3] S.T. Skou, F.S. Mair, M. Fortin, B. Guthrie, B.P. Nunes, J.J. Miranda, C.M. Boyd, S. Pati, S. Mtenga, S.M. Smith, Multimorbidity, Nat. Rev. Dis. Primers 8 (1) (2022) 48.

[4] A. Frølich, N. Ghith, M. Schiøtz, R. Jacobsen, A. Stockmarr, Multimorbidity, healthcare utilization and socioeconomic status: a register-based study in Denmark, PLoS One 14 (8) (2019) e0214183.

[5] S. Mercer, J. Furler, K. Moffat, D. Fischbacher-Smith, L. Sanci, Multimorbidity: Technical Series on Safer Primary Care, World Health Organization, 2016.

[6] C. Boyd, C.D. Smith, F.A. Masoudi, C.S. Blaum, J.A. Dodson, A.R. Green, A. Kelley, D. Matlock, J. Ouellet, M.W. Rich, et al., Decision making for older adults with multiple chronic conditions: executive summary for the American Geriatrics Society guiding principles on the care of older adults with multimorbidity, J. Am. Geriatr. Soc. 67 (4) (2019) 665–673.

[7] Academy of medical sciences, Multimorbidity: A Priority for Global Health Research, Academy of medical sciences, 2018.

[8] A. Hassaine, G. Salimi-Khorshidi, D. Canoy, K. Rahimi, Untangling the complexity of multimorbidity with machine learning, Mech. Ageing Dev. 190 (2020) 111325.

[9] L.T. Majnarić, F. Babič, S. O'Sullivan, A. Holzinger, AI and big data in healthcare: towards a more comprehensive research framework for multimorbidity, J. Clin. Med. 10 (4) (2021) 766.

[10] T.J. Loftus, B. Shickel, J.A. Balch, P.J. Tighe, K.L. Abbott, B. Fazzone, E.M. Anderson, J. Rozowsky, T. Ozrazgat-Baslanti, Y. Ren, et al., Phenotype clustering in health care: a narrative review for clinicians, Front. Artif. Intell. 5 (2022).

[11] D. Chushig-Muzo, C. Soguero-Ruiz, P. de Miguel-Bohoyo, I. Mora-Jiménez, Interpreting clinical latent representations using autoencoders and probabilistic models, Artif. Intell. Med. 122 (2021) 102211.

[12] D.T. Zemedikun, L.J. Gray, K. Khunti, M.J. Davies, N.N. Dhalwani, Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data, in: Mayo Clinic Proceedings, Vol. 93, Elsevier, 2018, pp. 857–866.

[13] C. Gadd, K. Nirantharakumar, C. Yau, mmVAE: multimorbidity clustering using relaxed Bernoulli β-variational autoencoders, in: Machine Learning for Health, PMLR, 2022, pp. 88–102.

[14] A.B. Jensen, P.L. Moseley, T.I. Oprea, S.G. Ellesøe, R. Eriksson, H. Schmock, P.B. Jensen, L.J. Jensen, S. Brunak, Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nat. Commun. 5 (1) (2014) 4022.

[15] K.W. Siah, C.H. Wong, J. Gupta, A.W. Lo, Multimorbidity and mortality: A data science perspective, J. Multimorbidity Comorbidity 12 (2022) 26335565221105431.

[16] G. Cezard, F. Sullivan, K. Keenan, Understanding multimorbidity trajectories in Scotland using sequence analysis, Sci. Rep. 12 (1) (2022) 16485.

[17] I.Y. Chen, R.G. Krishnan, D. Sontag, Clustering interval-censored time-series for disease phenotyping, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 6, 2022, pp. 6211–6221.

[18] D.P. Kingma, M. Welling, Auto-encoding variational bayes, 2013, arXiv preprint arXiv:1312.6114.

[19] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: International Conference on Machine Learning, PMLR, 2014, pp. 1278–1286.

[20] Y. Qin, M. van der Schaar, C. Lee, T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 3466–3492.

[21] D.L. Vetrano, A. Roso-Llorach, S. Fernández, M. Guisado-Clavero, C. Violán, G. Onder, L. Fratiglioni, A. Calderón-Larrañaga, A. Marengoni, Twelve-year clinical trajectories of multimorbidity in a population of older adults, Nature Commun. 11 (1) (2020) 3223.

[22] J. Zhao, Y. Zhang, D.J. Schlueter, P. Wu, V.E. Kerchberger, S.T. Rosenbloom, Q.S. Wells, Q. Feng, J.C. Denny, W.-Q. Wei, Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study, J. Biomed. Inform. 98 (2019) 103270.

[23] A. Hassaine, D. Canoy, J.R.A. Solares, Y. Zhu, S. Rao, Y. Li, M. Zottoli, K. Rahimi, G. Salimi-Khorshidi, Learning multimorbidity patterns from electronic health records using non-negative matrix factorisation, J. Biomed. Inform. 112 (2020) 103606.

[24] N. Dilokthanakul, P.A. Mediano, M. Garnelo, M.C. Lee, H. Salimbeni, K. Arulkumaran, M. Shanahan, Deep unsupervised clustering with gaussian mixture variational autoencoders, 2016, arXiv preprint arXiv:1611.02648.

[25] K.-L. Lim, X. Jiang, C. Yi, Deep clustering with variational autoencoder, IEEE Signal Process. Lett. 27 (2020) 231–235.

[26] A. Caciularu, J. Goldberger, An entangled mixture of variational autoencoders approach to deep clustering, Neurocomputing 529 (2023) 182–189.

[27] N. Jaques, S. Taylor, A. Sano, R. Picard, Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction, ACII, IEEE, 2017, pp. 202–208.

[28] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

[29] C. Violán, Q. Foguet-Boreu, S. Fernández-Bertolín, M. Guisado-Clavero, M. Cabrera-Bean, F. Formiga, J.M. Valderas, A. 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) (2019) e029594.

[30] G. Stafford, N. Villén, A. Roso-Llorach, A. Troncoso-Mariño, M. Monteagudo, C. Violán, Combined multimorbidity and polypharmacy patterns in the elderly: A cross-sectional study in primary health care, Int. J. Environ. Res. Public Health 18 (17) (2021) 9216.

[31] H. Hadipour, C. Liu, R. Davis, S.T. Cardona, P. Hu, Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means, BMC Bioinformatics 23 (4) (2022) 1–22.

[32] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.

[33] A. Stockmarr, A. Frølich, Clusters from chronic conditions in the danish adult population, PLOS ONE 19 (4) (2024) 1–24, http://dx.doi.org/10.1371/journal.pone.0302535.

[34] C.B. Pedersen, The danish civil registration system, Scand. J. Public Health 39 (7_suppl) (2011) 22–25.

[35] M. Schmidt, S.A.J. Schmidt, J.L. Sandegaard, V. Ehrenstein, L. Pedersen, H.T. Sørensen, The Danish National Patient Registry: a review of content, data quality, and research potential, Clin. Epidemiol. (2015) 449–490.

[36] O. Mors, G.P. Perto, P.B. Mortensen, The Danish psychiatric central research register, Scand. J. Public Health 39 (7_suppl) (2011) 54–57.

[37] H.W. Kildemoes, H.T. Sørensen, J. Hallas, The Danish national prescription registry, Scand. J. Public Health Suppl. (2011) 38–41.

[38] J. Sahl Andersen, N. De Fine Olivarius, A. Krasnik, The danish national health service register, Scand. J. Public Health 39 (7_suppl) (2011) 34–37.

[39] V.M. Jensen, A.W. Rasmussen, Danish education registers, Scand. J. Public Health 39 (7_suppl) (2011) 91–94.

[40] K. Helweg-Larsen, The danish register of causes of death, Scand. J. Public Health 39 (7_suppl) (2011) 26–29.

[41] K. Robinson, C. Lau, M. Jeppesen, A. Vind, C. Glümer, Kroniske Sygdomme-hvordan opgøres kroniske sygdomme? Region Hovedstaden, 2011.

[42] M.L. Schiøtz, A. Stockmarr, D. Høst, C. Glümer, A. Frølich, Social disparities in the prevalence of multimorbidity–A register-based population study, BMC Public Health 17 (1) (2017) 1–11.

[43] N.N. Holm, A. Frølich, O. Andersen, H.G. Juul-Larsen, A. Stockmarr, Longitudinal models for the progression of disease portfolios in a nationwide chronic heart disease population, Plos one 18 (4) (2023) e0284496.

[44] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, No. 14, Oakland, CA, USA, 1967, pp. 281–297.

[45] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[46] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[47] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (12) (2010).

[48] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, Beta-VAE: Learning basic visual concepts with a constrained variational framework, in: International Conference on Learning Representations, 2016.

[49] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, A. Lerchner, Early visual concept learning with unsupervised deep learning, 2016, arXiv preprint arXiv:1606.05579.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[51] L.N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, SPIE, 2019, pp. 369–386.

[52] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[53] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 807–814.

[54] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A k-means clustering algorithm, J. Royal Stat. Soc. C Appl. Stat. 28 (1) (1979) 100–108.

[55] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[56] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1985) 193–218.

[57] D.F. Crouse, On implementing 2D rectangular assignment algorithms, IEEE Trans. Aerosp. Electron. Syst. 52 (4) (2016) 1679–1696.

[58] S. van Oort, J.W. Beulens, A.J. van Ballegooijen, M.L. Handoko, S.C. Larsson, Modifiable lifestyle factors and heart failure: a Mendelian randomization study, Am. Heart J. 227 (2020) 64–73.

[59] A. Varbo, M. Benn, B.G. Nordestgaard, Remnant cholesterol as a cause of ischemic heart disease: evidence, definition, measurement, atherogenicity, high risk patients, and present and future treatment, Pharmacol. Ther. 141 (3) (2014) 358–367.

[60] B.A. Ference, H.N. Ginsberg, I. Graham, K.K. Ray, C.J. Packard, E. Bruckert, R.A. Hegele, R.M. Krauss, F.J. Raal, H. Schunkert, et al., Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel, Eur. Heart J. 38 (32) (2017) 2459–2472.

[61] A. Zajacova, E.M. Lawrence, The relationship between education and health: reducing disparities through a contextual approach, Annu. Rev. Public Health 39 (2018) 273–289.

[62] C.C. Low Wang, C.N. Hess, W.R. Hiatt, A.B. Goldfine, Clinical update: cardiovascular disease in diabetes mellitus: atherosclerotic cardiovascular disease and heart failure in type 2 diabetes mellitus–mechanisms, management, and clinical considerations, Circulation 133 (24) (2016) 2459–2502.

[63] M. Mehu, C.A. Narasimhulu, D.K. Singla, Inflammatory cells in atherosclerosis, Antioxidants 11 (2) (2022) 233.

[64] K. Barnett, S.W. Mercer, M. Norbury, G. Watt, S. Wyke, B. Guthrie, Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study, Lancet 380 (9836) (2012) 37–43.

[65] S.L. James, D. Abate, K.H. Abate, S.M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017, Lancet 392 (10159) (2018) 1789–1858.

[66] T. Stokes, Multimorbidity and clinical guidelines: problem or opportunity? New Zealand Med. J. (Online) 131 (1472) (2018) 7–9.

[67] S. Gautam, A. Boubekki, S. Hansen, S. Salahuddin, R. Jenssen, M. Höhne, M. Kampffmeyer, Protovae: A trustworthy self-explainable prototypical variational model, Adv. Neural Inf. Process. Syst. 35 (2022) 17940–17952.

[68] T.-Y. Cheng, M. Huertas-Company, C.J. Conselice, A. Aragon-Salamanca, B.E. Robertson, N. Ramachandra, Beyond the hubble sequence–exploring galaxy morphology with unsupervised machine learning, Mon. Not. R. Astron. Soc. 503 (3) (2021) 4446–4465.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[70] K. Iglay, H. Hannachi, P. Joseph Howie, J. Xu, X. Li, S.S. Engel, L.M. Moore, S. Rajpathak, Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus, Curr. Med. Res. Opin. 32 (7) (2016) 1243–1252.

[71] P. Hanlon, B.I. Nicholl, B.D. Jani, R. McQueenie, D. Lee, K.I. Gallacher, F.S. Mair, Examining patterns of multimorbidity, polypharmacy and risk of adverse drug reactions in chronic obstructive pulmonary disease: a cross-sectional UK biobank study, BMJ Open 8 (1) (2018) e018404.