mmVAE: multimorbidity clustering using Relaxed Bernoulli β-Variational Autoencoders

Charles Gadd* CWLGADD@GMAIL.COM
University of Oxford

Krishnarajah Nirantharakumar K.NIRANTHARAN@BHAM.AC.UK
University of Birmingham

Christopher Yau CHRISTOPHER.YAU@WRH.OX.AC.UK
University of Oxford, Alan Turing Institute, Health Data Research UK

Abstract

The prevalence of chronic disease multimorbidity is a significant and increasing challenge for health systems. In many cases, the occurrence of one chronic disease leads to the development of one or more other chronic conditions. This exerts a significant challenge in improving patient outcomes and is a growing challenge globally as average population age increases. Using electronic health record information to identify patterns of co-occurring conditions is seen as an unbiased means of understanding multimorbidity but most studies have adopted off-the-shelf algorithmic techniques that are not tailored for the application. We present a novel bespoke approach for multimorbidity clustering based on a highly customised version of a β-variational autoencoder. We incorporate the use of minimum entropy clustering to identify sparse, low-dimensional factored representations that link at a feature-level to the observed patient-level multimorbidity profiles. We demonstrate how the approach can be used to explore complex structure in a population-scale health data sets by examining data from a UK population of nearly 300,000 women in pregnancy suffering from multimorbidity.

* Corresponding author

1. Introduction

Many people in the most economically developed countries now live with multiple long-term chronic health conditions which can often result in significantly poorer quality of life (Whitty et al., 2020). Ageing is a significant contributory factor to these growing numbers with longer life expectancy leading to more age-related health conditions. However, younger age groups are also affected, often due to socioeconomic inequalities which contribute to raising levels of conditions such as obesity and mental health conditions but also unmet gender-related health issues.

This "multimorbidity" - the coexistence of two or more health conditions in the same individual - has significant impacts on the delivery of health services (Academy of medical sciences, 2018). Complex interactions between different health conditions can make treatment planning more complex (Hopman et al., 2016; Soley-Bori et al., 2021) and increases the likelihood of "polypharmacy" - prescribing multiple medications to the same patient - and the risks of adverse effects from interacting pharmaceutical agents (Osanlou et al., 2022).

Chronic health conditions often co-occur or *cluster* because of common risk factors. Only a fraction of the many thousands of

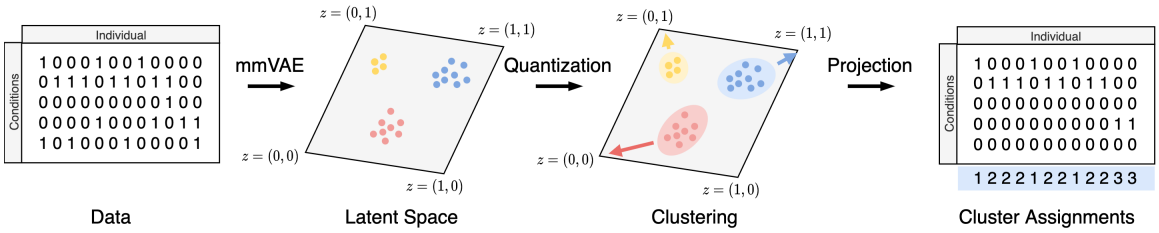

Figure 1: **Clustering derived by quantization.** The variational autoencoder maps observed binary data vectors into a latent space which is defined by an L-dimensional hypercube. In order to derive clusters, we map latent representations for each sample to the nearest corner of the hypercube and each corner defines a distinct cluster. The representation of each corner of the hypercube in terms of disease conditions can be obtained by projecting it back into the observation space through the decoder.

potential clusters are known and a canonical example is the association between diabetes and conditions of the skin, peripheral nerves, heart, eyes and brain. Mapping these clusters, deducing which are non-random, is crucial to uncover new disease mechanisms; develop treatments; and to reconfigure health services to better meet the needs of patients (Whitty and Watt, 2020).

With the availability of electronic medical records, systematic approaches have been developed to mine these records for multimorbidity cluster patterns. A standard approach is to examine cross-sectional data sets in which the health conditions of a large number of people are obtained and can be represented as a binary matrix. A model for this data must (i) scale to population-level data sets (ii) scale to large numbers of health conditions (iii) lead to a factored representation of overlapping groups of conditions.

Dimensionality reduction approaches including *Multiple Correspondence Analysis* (MCA) (Violán et al., 2019) or *Factor Analysis* (FA) (Poblador-Plou et al., 2014; Roso-Llorach et al., 2018) map to low-dimensional latent factors that can describe sets of associated conditions. Larsen et al. (2017) used *Latent Class Analysis* (LCA) to identify patterns of multimorbidity in a Danish population, but were severely limited to a small number of only 15 health conditions. Others, have performed similar studies, such as Violán et al. (2018) who used *k-means clustering*, and Hassaine et al. (2020) who used *non-negative matrix factorisation* (NMF). Probabilistic models such as Bernoulli mixture models (BMM) have also been adopted (Ng, 2015), whilst Markov-based models have been used to understand changes in multimorbidity over time (Bisquera et al., 2022).

Contributions

In this paper, we are specifically interested in developing a probabilistic clustering method (which we call `mmVAE`) that enables the exploration of multimorbidity patterns within *cross-sectional* population-level datasets. Taking as input a binary vector describing the set of conditions that an individual possesses, we map this to one of a number of possible groups or clusters that are described by underlying latent factors. Importantly, we want to obtain a model that

allows each patient to exhibit multiple latent factors which may, or may not, be shared between groups.

Our approach allows us to benefit from (i) non-linear dimensionality reduction via deep networks to handle large numbers of features, (ii) scalable computation through optimisation via end-to-end differentiability and programming environments designed for production-level systems (iii) whilst maintaining a probabilistic approach in this complex data setting. We make our code available at github.com/cwlgadd/mmVAE.

Clustering with multi-morbidity is complex, where many conflicting clusters can be concurrently concluded. Each of these can be correct in a secondary or tertiary care setting, where patients are grouped primarily by a clinician's speciality plus further morbidities. This approach is considered *comorbidity* clustering. For example, a cardiologist may take the interpretation of clusters centered around cardiovascular disease. In the context of Machine Learning and multimorbidity this can be interpreted as over-fitting.

Instead, in the primary care *multimorbidity* setting we are interested in a global clustering interpretation. In this case, we may then want to obtain the greatest reconstruction accuracy using the *fewest* bits of information. In this case we would be performing *minimum entropy clustering* and we would negate the entropy term in our objective.

Clustering methods based on information theory have proven to be successful (Sharma and Pemo, 2020; Li et al., 2004; Koltcov et al., 2019). They are based on the assumption that a cluster (or factor in our case), is one subset with the minimal possible degree of "disorder", and attempt to minimize the entropy of each. The motivation for this becomes clearer if we consider that the size of the entropy is inversely proportional to the degree of separation between the corresponding clusters/factors, so the classification using the feature with the smallest entropy (that is, the feature with the least uncertainty) is the best (Jia et al., 2022).

2. Methodology

Let $\mathbf{x} = [x_1, \dots, x_D] \in \{0,1\}^D$ be a vector which represents a subject's diagnoses, where 1 denotes a positive diagnosis for each of the D possible conditions and let $\mathbf{z} \in \{0,1\}^L$ denote a latent embedding in an L-dimensional binary vector space where typically $L \ll D$.

The goal is to learn a generative autoencoder model, whose joint distribution is given by $p(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ and $\boldsymbol{\theta}$ is the set of bias and weight parameters of a *decoder* neural network. From this the posterior distribution $p(\mathbf{z}|\mathbf{x})$ can be obtained which provides a mechanism for clustering patients morbidity patterns as 2^D possible observed morbidity patterns must be reduced to 2^L in the latent space. Given a training set $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$, the model is trained by maximising the log marginal likelihood with respect to the parameters,

$$\log p\left(\mathbf{X}\right) = \sum_{i=1}^{N} \log \int p\left(\mathbf{x}_i|\mathbf{z}_i\right) p\left(\mathbf{z}_i\right) d\mathbf{z}_i.$$

Variational autoencoders. This marginalisation over the latent variables is often intractable. To address this Kingma and Welling (2013) introduced a parametric, amortized variational distribution $q_{\boldsymbol{\phi}}\left(\mathbf{z}\right)$ called an *encoder* to approximate the true but intractable posterior distribution $p\left(\mathbf{z}|\mathbf{x}\right)$. Using this approximation we are then able to obtain a variational evidence lower bound (ELBO) of our marginal likelihood objective which we can choose to optimise with gradient based approaches

$$\log p\left(\mathbf{x}\right) \geq \sum_{i=1}^{N} \Bigg[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i)} \left[\log p_{\boldsymbol{\theta}}\left(\mathbf{x}_i|\mathbf{z}_i\right)\right] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}\left(\mathbf{z}_i\right) || p\left(\mathbf{z}_i\right)\right) \Bigg].$$

Unbiased estimates of the gradient of the first term with respect to $\boldsymbol{\theta}$ are easy to obtain through Monte Carlo sampling, but gradients with respect to the variational parameters $\boldsymbol{\phi}$ are more challenging to obtain and are often achieved through a reparameterisation trick, in which the stochastic sampling routine for $\boldsymbol{z}$ is replaced by a deterministic and differentiable affine function $\mathbf{z} = f(\boldsymbol{\epsilon}, \mathbf{x})$ of an auxiliary random variable $\boldsymbol{\epsilon}$ with a parameter-independent base distribution $q(\boldsymbol{\epsilon})$.

The Relaxed Bernoulli distribution. This reparameterisation trick cannot be directly applied to discrete random variables as there is no affine function which can transform the base distribution to a discrete distribution. The Concrete distribution (Maddison et al., 2017; Jang et al., 2016), resolves this issue with relaxation of the categorical distribution based on the Gumbel-Softmax trick, where sampling of a discrete random variables is refactored. The Relaxed Bernoulli distribution is a binary special case and is a continuous relaxation of the Bernoulli distribution with support on the unit interval $(0, 1)$. Using this, the sampling procedure for random variable $B \sim \text{RelaxedBernoulli}(\alpha, \lambda)$ can be reparameterised through the formulation

$$l = \log(\alpha) + \log(u) + \log(1-u),$$
$$B = \frac{1}{1+\exp(-l/\lambda)},$$

where $u \sim \text{Uniform}(0, 1)$, $\alpha \in (0, \infty)$ is a location parameter and $\lambda \in (0, \infty)$ is an annealed temperature parameter that controls the degree of the approximation. As the temperature $\lambda \to 0$ the random variable B converges in distribution, such that $B \dot{\sim} \text{Bernoulli}(\mu)$, with $\mu = \alpha/1 + \alpha$. We now have

$$q_{\boldsymbol{\phi}}(\mathbf{z}) = \prod_{l=1}^{L} \text{RelaxedBernoulli}(z_l|\alpha_l; \boldsymbol{\phi}, \lambda),$$

where the temperature λ is annealed over an exponentially decayed training schedule similar to that proposed in Jang et al. (2016).

We have assumed independence between latent dimensions which leads to a mixture model latent construction. Consequently any inferred latent medical factors (e.g. cardiovascular or mental health related diseases) are assumed independent. This assumption is often (implicitly) made, but can be removed if desired by considering the full covariance (Wang and Yin, 2020).

Entropy encoding. While standard VAEs use multivariate Gaussian distributions as a prior for the latent variables $\mathbf{z}$, we adopt a factorised uniform prior instead

$$p(\mathbf{z}) = \prod_{l=1}^{L} \text{Uniform}(z_l; 0, 1).$$

This choice of uniform prior leads to our KL divergence term taking the form of the entropy of $q_\phi(\mathbf{z})$

$$D_{\text{KL}}(q||p) = \sum_{l=1}^{L} \sum_{z_l \in \{0,1\}} q(z_l) \log q(z_l).$$

As the entropy of the Relaxed Bernoulli distribution does not exist in closed form, we take the limit in which $\lambda = 0$, and approximate using the standard Bernoulli distribution, $D_{\text{KL}}(q||p) = \sum_{l=1}^{L} \left[\mu_l \log \mu_l + (1-\mu_l)\log(1-\mu_l)\right]$, where μ_l are the probabilities of the relaxed encoder Bernoulli distribution in Eq. (2). We can now observe that the divergence of our ELBO takes the form of the sum of the negative *binary entropy*, $-\mathbf{H}_b(q)$, of our variationally encoded samples, summed independently across each latent factor. Derivation details are given in Appendix A.

Maximising over the evidence lower bound then equates to maximising a reconstruction term and the binary entropy. The contribution of this entropy term is controlled through

a normalised scalar parameter $\hat{\beta} = \frac{\beta D}{L}$. A higher positive weight then leads to finding the latent representation with greatest uncertainty, which typically has the upshot of leading to greater generalisation (Higgins et al., 2017; Burgess et al., 2018). As the greatest separation minimises the entropy, we choose β to be negative leading to *minimal entropy clustering* within factors. The marginal likelihood bound we optimise over is then given by

$$\begin{aligned} &\log p\left(\mathbf{x}\right) \geq \mathcal{L}\left(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}\right) \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}\left(\mathbf{x}|\mathbf{z}\right)\right] + \hat{\beta} \sum_{l=1}^{L} \mathbf{H}_b\left(q\right). \end{aligned} \tag{1}$$

Encoder and decoder architectures. We use fully connected layers, with the L_0 sparsification of Louizos et al. (2017) which automatically prunes our model. Often with variational autoencoders we find that complex architectures are not needed (Flenner and Hunter, 2017), and dependent on the scale of the divergence only one of the encoder (for autoencoding) or decoder (for autodecoding) is required to be more complex. Regularisation in our autoencoder allows automatic scaling with increased complexity without overfitting. We may also have considered the effect of dropout (Srivastava et al., 2014), but this approach prevents complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors. Rather, we hope to specifically leverage these co-adaptions. *Batch normalisation* is used following each fully connected layer. *ReLU activation functions* are used on hidden layers and the *sigmoid* link activation function is used in the final decoder layer.

Clustering by discretisation. In order to recover the binary latent representation from the Relaxed Bernoulli-based approximate inference, we map the expected values of $\mathbb{E}_{q(\mathbf{z}_i)}[\mathbf{z}_i]$ to the nearest vertex of an L-dimensional latent hypercube. This is visually depicted in Fig. 1. The effect of this is to create clusters of individuals with similar representations and therefore health conditions. The representation of each hypercube corner in terms of observed health conditions can be obtained by mapping each corner through the trained decoder. Therefore each (quantised) latent dimension can be associated with a set of health conditions allowing each latent dimension to be considered as interpretable *factors* - subsets of recurrent morbidities - that combine to explain each individuals morbidity profile.

Cluster and factor descriptions. We will assign a qualitative interpretation to our clustering and latent factors by examining the relative proportion of conditions possessed by individuals associated with these clusters/factors versus the remaining population. For clusters, this is given as the odds of a health condition occurring given a patient belongs to a cluster, versus the odds without. For factors, this is given by the odds of a health condition occurring given a patient belongs to a factor, versus the odds without. For a condition d, factor l or cluster c (defined as a unique combination of factors), these are respectively given by

$$\begin{aligned} \mathrm{OR}_{d|c} &= \frac{\mathbb{P}\left(\text{condition } d|\text{cluster } = c\right)}{\mathbb{P}\left(\text{condition } d|\text{cluster } \neq c\right)} \\ \mathrm{OR}_{d|l} &= \frac{\mathbb{P}\left(\text{condition } d|\hat{z}_l = 1\right)}{\mathbb{P}\left(\text{condition } d|\hat{z}_l = 0\right)}, \end{aligned} \tag{2}$$

where $\hat{z}_i$ denotes a quantised latent element. For visual clarity, we graphically present truncated odds ratios to not exceed 5 to be interpretable for the cases where the denominators are low or zero valued, and truncate those below 1 to zero.

3. Multi-morbidity in pregnancy

Study and data description. We designed a series of experiments based on curated subsets

from the MuM-PreDiCT study Lee et al. (2022). We extracted a sub-population of 513,631 individuals who were diagnosed with one or more of 79 different health conditions before or during the term of the pregnancy. Of these cases, 41,971 unique health condition combinations of a potential 2^{79} were observed. We retained only patients who fulfilled the definition of multimorbidity - diagnoses of two or more health conditions - and for patients who had multiple recorded pregnancies we kept only the last record. This left a final data set containing 290,658 individuals and 41,893 unique health condition combinations. Table 1 presents the prevalence of a number of selected health conditions, and the number of patients exhibiting different levels of multimorbidity. The top four most common health conditions were depression, anxiety, allergic rhinoconjunctivitis, and asthma. The full table is presented in the Supplementary Material.

Reproducibility. Unless otherwise specified we present results for $\beta = -0.4$, but results for a range of values were explored. The adjusted mutual information across 10 seeds is given in Fig. 6 and further plots are included in the Appendix and Supplementary Material. Throughout we anneal the temperature parameter of the Relaxed Bernoulli distribution λ from 4 to 0.4 and use $L = 16$ which leads to the capacity to express $2^{16} = 65,536$ unique latent codes. This choice then theoretically allows every one of the 41,893 unique combination of observed patient profiles to be its own cluster, thus a choice of $L = 16$ was deemed to be sufficiently (and likely overly-) expressive.

Subset analysis on three unrelated conditions. We first examined a subset ($N = 106,593$) consisting of patients who had a diagnosis of (i) asthma, (ii) cancer (all types) or (iii) female infertility. The prevalence and number of co-morbidities of each of these conditions is shown in Table 1.

As this is real-world observational data, there are *no* ground-truth labels, but our subset analysis on a curated dataset will inform the parameters necessary for the exploratory analysis of the larger MuM-PreDiCT data set. For this subset, it might be expected that any analysis would yield three distinct subgroups due to the clinical orthogonality of the selected conditions. However, the presence of multimorbidity in these individuals leads to far more complex sub-structures.

Figure 2 shows the first 24 most prevalent clusters (consisting of 85.1% of the analysis population) identified by applying mmVAE to the data. The ten most prevalent clusters involve 63% of the analysis population and contain a distinct cluster associated with infertility (cluster 4) which occurs without asthma but is linked to range of conditions including pituitary disorders and gynaecological conditions (polycystic ovary syndrome, endometriosis and leiomyoma) which have all been linked to female infertility. The remaining nine clusters exhibit some association with asthma but are distinguished by associations with depression (clusters 3, 5, 7, 10), human immunodeficiency virus (clusters 2, 4), allergic rhinoconjunctivitis (clusters 1, 7, 9, 10), migraine (clusters 6, 9), Turner's syndrome (and organ transplant*) (clusters 7*, 9), and anxiety (clusters 5, 6, 10). Cancer is not represented as a distinct cluster which is a symptom of the fact of its low population prevalence and the clustering is dominated by combinations of the most prevalent health condition.

We next examined the latent factors that comprise each identified cluster. Figure 4 shows that there is a sparse assignment of factors to clusters when minimising entropy ($\beta = -0.4$) in contrast to when entropy is being maximised (Appendix Figure 8). Figure 3 illustrates the health conditions associated

Primary condition	Prevalence ($N = 290,658$)	Total co-morbidities				
		1	2	3	4	5+
Depression	133,060 (45.8%)	41,900	37,532	25,283	14,567	13,778
Anxiety	97,239 (33.5%)	27,619	27,039	19,273	11,676	11,632
Allergic Rhinoconjunctivitis	92,340 (31.8%)	33,987	24,924	15,739	8,836	8854
Asthma	85,315 (29.4%)	30,295	22,920	14,868	8,498	8,734
Female infertility	24,121 (8.3%)	8,889	6,465	4,108	2,280	2,379
Cancer	2,890 (1.0%)	957	791	498	283	361

Table 1: The total number of patients with health conditions of most interest, and the total number of patients with different numbers of co-morbidities. A table for all 79 health conditions can be found in the Supplementary Material.

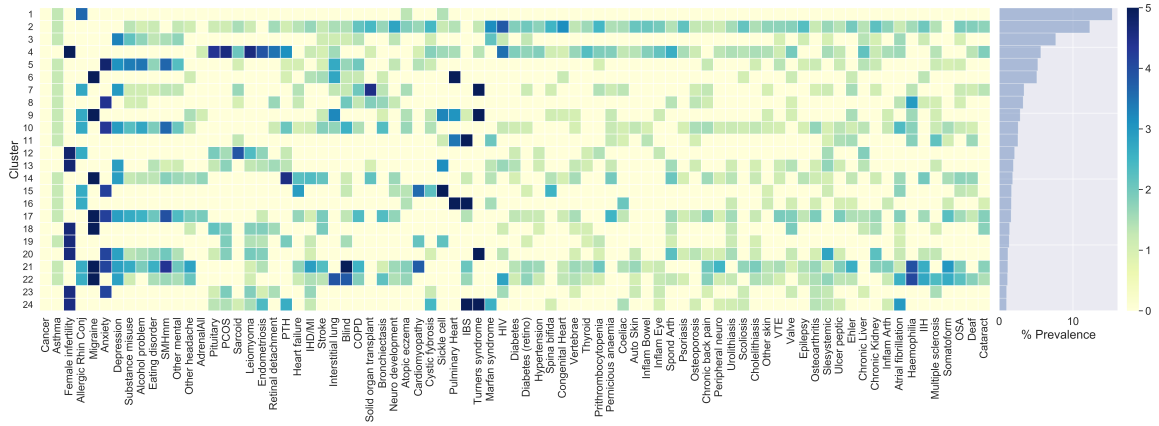

Figure 2: Clustering odds ratio on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions, and the percentage of individuals in each cluster. Clusters with fewer than $< 1\%$ of the samples are not shown.

with each factor. Note that the latent factors 5, 8, 15 and 16 are associated with cancers which was not observable directly from the cluster structure in Figure 2. Latent factor 8 is utilised by most of the ten most prevalent clusters indicating that cancers occur sporadically and independently of other health conditions. The exception of cluster 4 which is linked to infertility which instead uses factors 5 and 16 (these are duplicated latent factors) to explain the presence of cancers within the cluster. Both factors possess an explicit link between infertility and cancer.

We note that our decoder means that each cluster is a non-linear combination of factors but we do not constrain the decoder to be additive only combinations. Additive structure, whilst interpretable, would encourage a basis-like behaviour in which latent factors comprise of single, pairs, triplets, etc of conditions leading to a non-sparse clustering structure.

Our analysis suggests that `mmVAE` is therefore able to find patterns and associations of health conditions that through retrospective interpretation are consistent with exist-

ing clinical knowledge. The motivation for automated learning is exemplified by the fact that though we seemingly created a dataset in which there were *seemingly* three obvious clusters, the presence of complex multimorbidity leads to alternate explanations of the data.

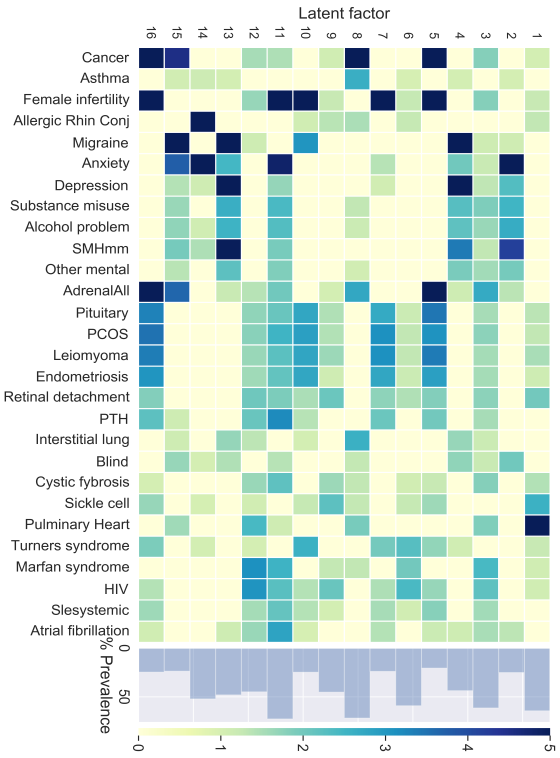

Figure 3: Factor odds ratio on a cohort with (i) asthma (ii) cancer and (iii) female infertility, and the percentage of individuals with each factor. Conditions with no odds ratio above 2 are filtered.

Encoding out-of-distribution. We next tested the out-of-distribution generalisability of the trained model from the previous experiment by encoding on a test cohort which comprises of patients who have depression. We exclude those who exhibited co-morbidities with asthma, cancer and female infertility,

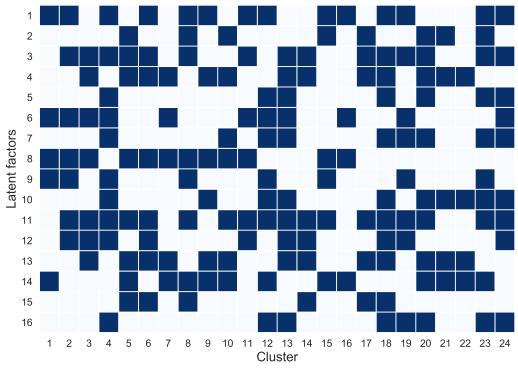

Figure 4: Cluster-factor association matrix.

that were used in the training process. This leads to a test set of 94,905 samples, each composed of previously unseen combinations of health conditions. We map each of these through the encoder and record each patient's factors and consequent cluster.

The histogram of factor and cluster prevalence (Figure 5) shows that all depression individuals were projected on to factors 5, 13 and 16. Factor 13 is associated with a range of mental health conditions including depression while factors 5 and 16 were exclusive of these conditions. As the decoder uses both positive and negative contributions from latent factors, the relevance of these factors is clear in that combinations of these can modulate the cluster-specific probabilities for depression and mental health conditions. The depression cohort is mapped to 22 clusters, none of which were common clusters observed in the training set, and 13 were seen rarely in the training cohort. The remaining 9 were novel combinations of latent factors.

Choice of β. We may consider how independently and identically trained models differ by considering the adjusted mutual information of clustering, averaged equally over 10 independent runs, shown in Fig. 6. This metric measures how much information is shared between the clustering of independently trained models. We observe that as our model shifts

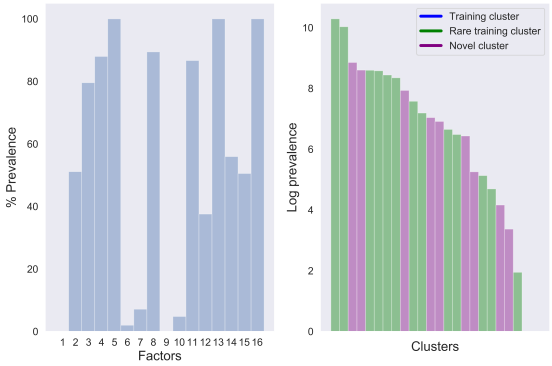

Figure 5: Test factor and cluster prevalence.

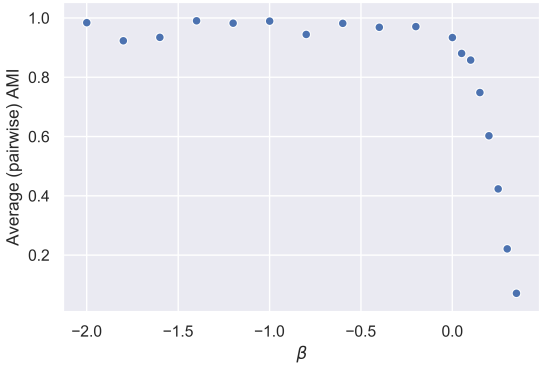

Figure 6: Adjusted mutual information averaged from 10 independent trials.

from minimal ($\beta < 0$) to maximal entropy ($\beta > 0$), the adjusted mutual information sharply decreases. This is likely a consequence of the model over-fitting to isolated clinical perspectives.

Analysis of full `MuM-PreDiCT` *data.* We then applied `mmVAE` to the full `MuM-PreDiCT` study data set of 290,658 individuals. Figure 7 shows a t-SNE visualisation of a subset of 10,000 individuals (Appendix B) which illustrates the lack of any clear sub-structure and the motivation for our bespoke approach which focuses on detailed feature-level modelling. Clustering and latent factor profiles are given in Figures 8 and 9 (Appendix B). On the complete `MuM-PreDiCT` data, the largest 24 clusters now represent 66.6% of the population. As before, while infertility is associated with a few number of clusters (primarily cluster 15), asthma and cancer are spread over a number of clusters. Our motivation for first examining a curated subset previously is demonstrated by the emergence of *super-clusters* (clusters 2, 4-6, 8-10) which are associated with a large number of conditions. This is exemplified by cluster 2 which is strongly associated with increased odds of heart failure which can arise due to a number of factors such as complications with diabetes, spina bifida, HIV as well as being the end point of cardiovascular disease - all are prominent in cluster 2 with increased odds. While the most prevalent cluster 1 is largely associated with anxiety and depression. Space precludes a deeper discussion of these findings but the structures identified by `mmVAE` have been reviewed by clinically qualified associates for applicability.

4. Conclusion

We have presented a novel framework, `mmVAE`, for multimorbidity clustering based on minimising information entropy in the β-variational autoencoder. We demonstrated the use of the algorithm for a novel exploratory cross-sectional analysis of multimorbidity in pregnancy using a population-scale UK electronic health dataset. Our approach is a bespoke approach that is specifically designed for clustering of binary data of this type where clusters are assumed to share common binary latent factors. The implementation using modern deep learning framework allows for scalable implementations and we have had the rare opportunity in this study to have been able to apply a state-of-the-art approach to a sensitive, protected health data set.

Acknowledgements

We thank our clinically qualified colleagues of the `MuM-PreDiCT` team for reviewing the algorithmic outputs for clinical applicability.

Ethics and consent statement

Data access has been granted by the CPRD Independent Scientific Advisory Committee (reference: 20_181R). As the study data are de-identified, consent is not required.

Funding statement

This work is funded by the Strategic Priority Fund "Tackling multimorbidity at scale" programme (grant number MR/W014432/1) delivered by the Medical Research Council and the National Institute for Health Research (NIHR) in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council. The views expressed are those of the author and not necessarily those of the funders, the NIHR or the UK Department of Health and Social Care. The funders had no role in study design, decision to publish, or preparation of the manuscript. CY and CG are also supported by an EPSRC Turing AI Acceleration Fellowship (Grant Ref: EP/V023233/1).

References

Academy of medical sciences. *Multimorbidity: a priority for global health research*. Academy of medical sciences, 2018.

Alessandra Bisquera, Ellie Bragan Turner, Lesedi Ledwaba-Chapman, Rupert Dunbar-Rees, Nasrin Hafezparast, Martin Gulliford, Stevo Durbaba, Marina Soley-Bori, Julia Fox-Rushby, Hiten Dodhia, et al. Inequalities in developing multimorbidity over time: A population-based cohort study from an urban, multi-ethnic borough in the united kingdom. *The Lancet Regional Health-Europe*, 12:100247, 2022.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Jennifer Flenner and Blake Hunter. A deep non-negative matrix factorization neural network. *Semantic Scholar*, 2017.

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, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Petra Hopman, Simone R De Bruin, Maria João Forjaz, Carmen Rodriguez-Blazquez, Giuseppe Tonnara, Lidwien C Lemmens, Graziano Onder, Caroline A Baan, and Mieke Rijken. Effectiveness of comprehensive care programs for patients with multiple chronic conditions or frailty: a systematic literature review. *Health policy*, 120(7):818–832, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduc-

tion: a review. *Complex & Intelligent Systems*, pages 1–31, 2022.

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

Sergei Koltcov, Vera Ignatenko, and Sergei Pashakhin. Fast tuning of topic models: an application of rényi entropy and renormalization theory. *Multidisciplinary Digital Publishing Institute Proceedings*, 46(1):5, 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.

Siang Ing Lee, Amaya Azcoaga-Lorenzo, Utkarsh Agrawal, Jonathan I Kennedy, Adeniyi Francis Fagbamigbe, Holly Hope, Anuradhaa Subramanian, Astha Anand, Beck Taylor, Catherine Nelson-Piercy, et al. Epidemiology of pre-existing multimorbidity in pregnant women in the uk in 2018: a population-based cross-sectional study. *BMC pregnancy and childbirth*, 22(1):1–15, 2022.

Haifeng Li, Keshu Zhang, and Tao Jiang. Minimum entropy clustering and applications to gene expression analysis. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, pages 142–151. IEEE, 2004.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. 2017.

SK Ng. A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among australians. *Statistics in Medicine*, 34(26):3444–3460, 2015.

Rostam Osanlou, Lauren Walker, Dyfrig A Hughes, Girvan Burnside, and Munir Pirmohamed. Adverse drug reactions, multimorbidity and polypharmacy: a prospective analysis of 1 month of medical admissions. *BMJ open*, 12(7):e055551, 2022.

Beatriz Poblador-Plou, Amaia Calderón-Larrañaga, Javier Marta-Moreno, Jorge Hancco-Saavedra, Antoni Sicras-Mainar, Michael Soljak, and Alexandra Prados-Torres. Comorbidity of dementia: a cross-sectional study of primary care older patients. *BMC psychiatry*, 14(1):1–8, 2014.

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.

Shachi Sharma and Sonam Pemo. Performance analysis of various entropy measures in categorical data clustering. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 592–595. IEEE, 2020.

Marina Soley-Bori, Mark Ashworth, Alessandra Bisquera, Hiten Dodhia, Rebecca Lynch, Yanzhong Wang, and Julia Fox-Rushby. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the uk literature. *British Journal of General Practice*, 71(702):e39–e46, 2021.

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.

Concepción Violán, Albert Roso-Llorach, Quintí Foguet-Boreu, Marina Guisado-Clavero, Mariona Pons-Vigués, Enriqueta Pujol-Ribera, and Jose M Valderas. Multimorbidity patterns with k-means nonhierarchical cluster analysis. *BMC family practice*, 19(1):1–11, 2018.

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.

Xi Wang and Junming Yin. Relaxed multivariate bernoulli distribution and its applications to deep generative models. In *Conference on Uncertainty in Artificial Intelligence*, pages 500–509. PMLR, 2020.

Christopher JM Whitty and Fiona M Watt. Map clusters of diseases to tackle multimorbidity, 2020.

Christopher JM Whitty, Carrie MacEwen, Andrew Goddard, Derek Alderson, Martin Marshall, Catherine Calderwood, Frank Atherton, Michael McBride, John Atherton, Helen Stokes-Lampard, et al. Rising to the challenge of multimorbidity, 2020.

Appendix A. Evidence lower bound derivations

For our Relaxed Bernoulli β-VAE, we chose to approximate the KL divergence term in the ELBO by computing the KL divergence between the discretization of the relaxed posterior and the discrete uniform prior, which corresponds to Eq. (22) in Appendix C of Maddison et al. (2017) and was also used in the official implementation of categorical VAE with the Gumbel-Softmax estimator (Jang et al., 2016). The divergence is therefore between the choice of factored prior and

$$q(\mathbf{z}) = \prod_{l=1}^{L} \text{RelaxedBernoulli}(z_l; \boldsymbol{\phi}, \lambda = 0).$$

Under discretisation of the relaxed Bernoulli distribution, the divergence follows in the following steps.

$$\begin{aligned} &-D_{\text{KL}}(q||p) \\ &\quad = -\int_{\mathbf{z}\in\mathbb{R}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} \\ &\quad = -\int_{\mathbf{z}\in\mathbb{R}} \prod_{i=1}^{L} q(z_i) \log \prod_{j=1}^{L} \frac{q(z_j)}{p(z_j)} d\mathbf{z}. \end{aligned}$$

Factored Bernoulli prior

$$p(\mathbf{z}) = \prod_{l=1}^{L} \text{Bernoulli}(z_l).$$

For completeness we include the derivation for a Bernoulli prior, which also leads to an entropy interpretation.

$$\begin{aligned} &-D_{\text{KL}}(q||p) \\ &\quad = -\int_{\mathbf{z}\in[0,1]} \sum_{j=1}^{L} \prod_{i=1}^{L} q(z_i) \log \frac{q(z_j)}{p(z_j)} d\mathbf{z} \\ &\quad = -\sum_{j=1}^{L} \sum_{z} \prod_{i=1}^{L} q(z_i) \log \frac{q(z_j)}{p(z_j)}. \end{aligned}$$

Let $z_{i\backslash j}$ denote the vector that results from taking out the j^{th} component of z, and $q_{z_{i\backslash j}}(\cdot)$ its corresponding factored distribution. This allows us to re-write this as

$$\begin{aligned}
&-D_{\mathrm{KL}}(q||p) \\
&\quad= -\sum_{j=1}^{L}\sum_{z_{i\backslash j},z_j}\prod_{i=1}^{L} q(z_i)\log\frac{q(z_j)}{p(z_j)} \\
&\quad= -\sum_{j=1}^{L}\sum_{z_{i\backslash j},z_j} q\left(\mathbf{z}_{i\backslash j}\right) q(z_j)\log\frac{q(z_j)}{p(z_j)} \\
&\quad= -\sum_{j=1}^{L}\sum_{z_{i\backslash j}} q\left(\mathbf{z}_{i\backslash j}\right)\overbrace{\sum_{z_j} q(z_j)}^{1}\log\frac{q(z_j)}{p(z_j)}.
\end{aligned}$$

It then follows (under discretisation of the relaxed distribution) that the negative divergence takes the form

$$\begin{aligned}
&-D_{\mathrm{KL}}(q||p) \\
&\quad= -\sum_{j=1}^{L}\sum_{z_j} q(z_j)\log\frac{q(z_j)}{p(z_j)} \\
&\quad= -\sum_{j=1}^{L} q_j\log\frac{q_j}{p_j} + (1-q_j)\log\frac{1-q_j}{1-p_j},
\end{aligned}$$

where q_j and p_j are the probability of a latent factor dimension being 'turned on'

$$\begin{aligned}
&-D_{\mathrm{KL}}(q||p) \\
&\quad= -\sum_{j=1}^{L}\overbrace{\sum_{z_j} q(z_j)\log q(z_j)}^{-\mathbf{H}_b(q)} \\
&\qquad\qquad - q(z_j)\log p(z_j).
\end{aligned}$$

Factored uniform prior

$$p(\mathbf{z}) = \prod_{l=1}^{L}\mathrm{Uniform}(z_l; 0, 1).$$

Using a combination of logarithmic rules, the pdf of the uniform prior $\mathbb{P}(\mathbf{z}=v) = \frac{1}{1-0}\mathbb{I}_{(0,1)}(v)$, defined over the unit interval latent support we obtain

$$\begin{aligned}
&-D_{\mathrm{KL}}(q||p) \\
&\quad= -\int_{\mathbb{R}}\sum_{j=1}^{L}\prod_{i=1}^{L} q(z_i)\log\frac{q(z_j)}{\mathbb{I}_{[0,1]}(z)}d\mathbf{z} \\
&\quad= -\int_{[0,1]}\sum_{j=1}^{L}\prod_{i=1}^{L} q(z_i)\log q(z_j)\,d\mathbf{z} \\
&\quad= -\sum_{j=1}^{L}\sum_{z}\prod_{i=1}^{L} q(z_i)\log q(z_j).
\end{aligned}$$

Using the same trick as before

$$\begin{aligned}
&-D_{\mathrm{KL}}(q||p) \\
&\quad= -\sum_{j=1}^{L}\sum_{z_{i\backslash j}} q\left(\mathbf{z}_{i\backslash j}\right)\overbrace{\sum_{z_j} q(z_j)}^{1}\log q(z_j) \\
&\quad= -\sum_{j=1}^{L}\sum_{z_j} q(z_j)\log q(z_j) \\
&\quad= \sum_{j=1}^{L}\mathbf{H}_b(q).
\end{aligned}$$

The constant β is then applied and normalised, under the framework outlined in Higgins et al. (2017) and Burgess et al. (2018).

Appendix B. Full Data Analysis

In this section we provide the output of our model on the full CPRD dataset, which was removed in the main text for brevity.

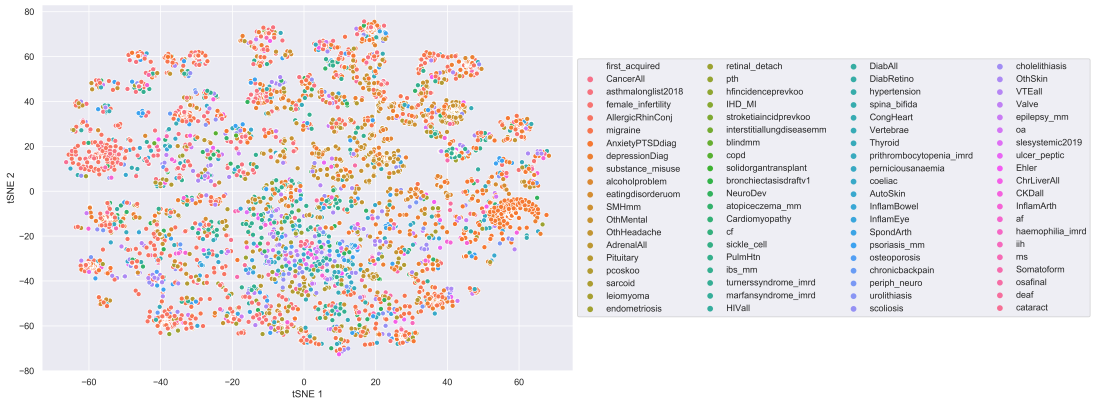

Figure 7: `MuM-PreDiCT` t-SNE visualisation. Individual patient points are colored by the first condition acquired.

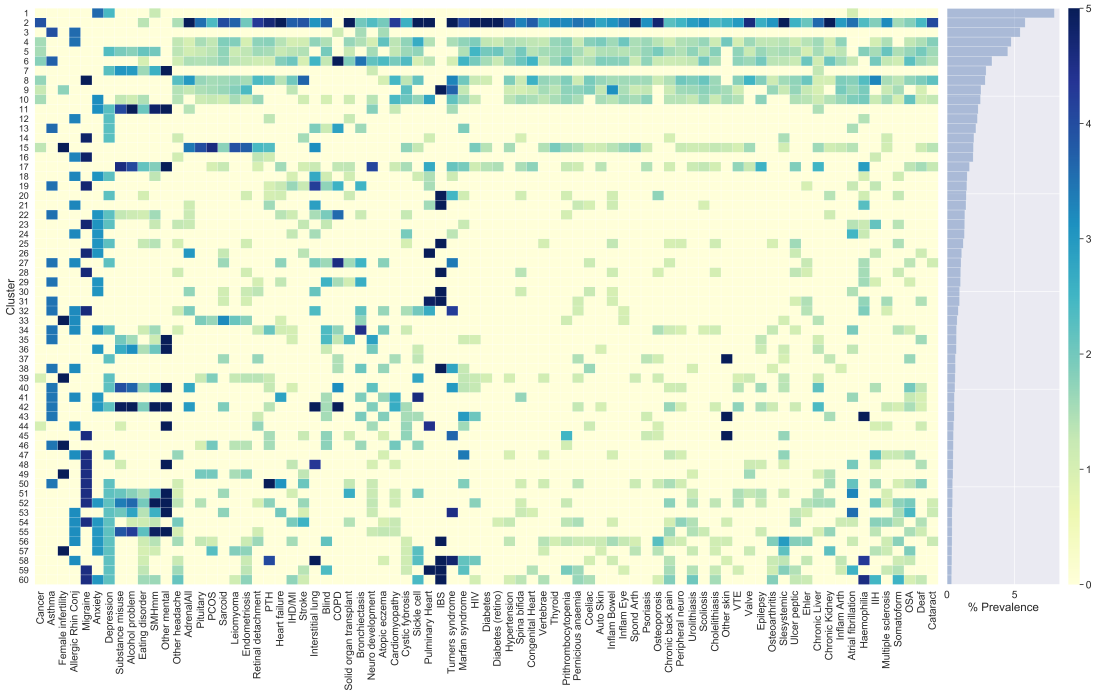

Figure 8: `MuM-PreDiCT` clusters.

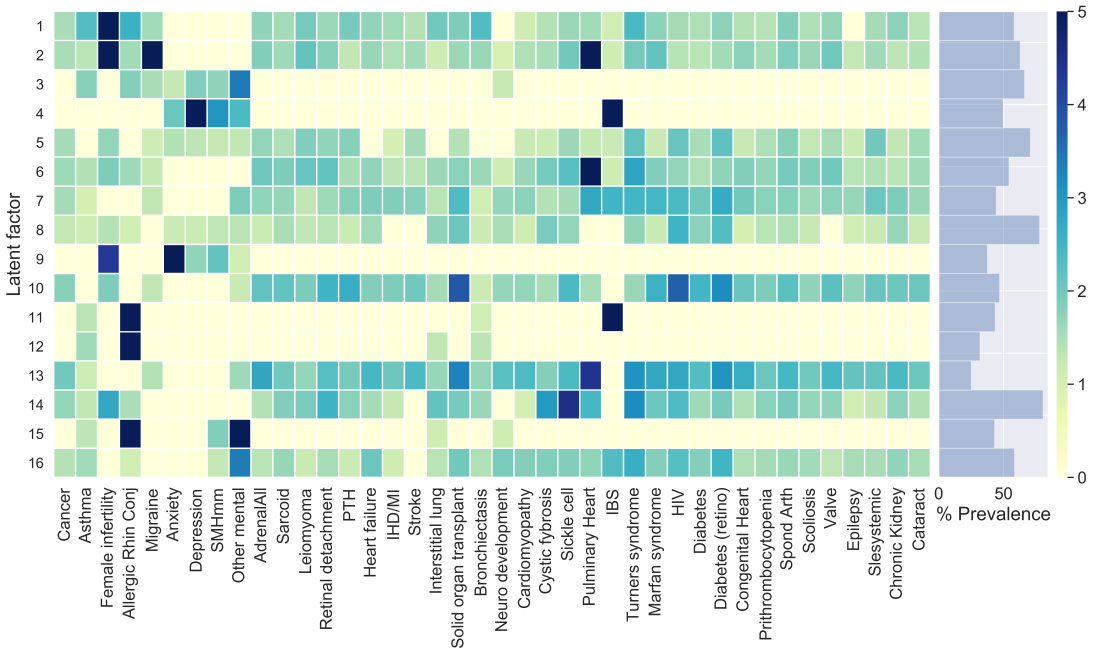

Figure 9: `MuM-PreDiCT` latent factors.