Feature Allocation Approach for Multimorbidity Trajectory Modelling

Woojung Kim WOOJUNG.KIM@WARWICK.AC.UK
University of Warwick, The Alan Turing Institute

Paul A. Jenkins P.JENKINS@WARWICK.AC.UK
University of Warwick, The Alan Turing Institute

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

Abstract

A multimorbidity trajectory charts the time-dependent acquisition of disease conditions in an individual. This is important for understanding and managing patients who have a complex array of multiple chronic conditions, particularly later in life. We construct a novel probabilistic generative model for multimorbidity acquisition within a Bayesian framework of latent feature allocation, which allows an individual's morbidity profile to be driven by multiple latent factors and allows the modelling of age-dependent multimorbidity trajectories. We demonstrate the utility of our model in applications to both simulated data and disease event data from patient electronic health records. In each setting, we show our model can reconstruct clinically meaningful latent multimorbidity patterns and their age-dependent prevalence trajectories.

Keywords: Multimorbidity Analysis, Bayesian Feature Allocation Model, Markov Chain Monte Carlo

1. Introduction

Multimorbidity refers to individuals who have two or more medical conditions simultaneously and can also be referred to as "multiple (long-term) health conditions". Multimorbidity presents a major challenge for clinicians due to the interacting effects of different conditions and the treatment approaches that may be employed. It is often not clear what the optimal clinical management approach should be for any given patient. The prevalence of multimorbidity is increasing globally due to factors such as ageing populations and socioeconomic inequalities.

While some health conditions occur together by chance, others are non-random due to a background of common genetic or environmental pathways. As a consequence, there is increasing recognition that multimorbidity cannot be thought of as a random assortment of individual conditions but as a series of predictable time-evolving 'groups' of conditions within individuals. This has led to an increasing interest in using large-scale population-scale data sets to obtain evidence for recurrent multimorbidity patterns.

Computational phenotyping techniques (e.g. Hassaine et al., 2020) can be used when there is access to full electronic health records (EHRs). If the diagnosis times for conditions acquired by each individual are available, it may be possible to use matrix or tensor factorisation approaches to determine the latent factors corresponding to groups of associated conditions as well as temporal factors describing their evolution. Alter-

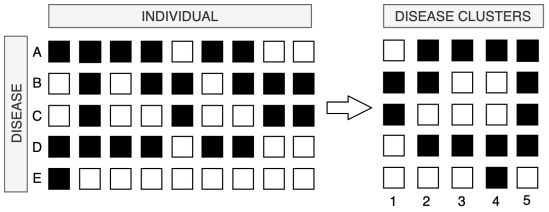

Figure 1: **Multimorbidity data**. Data can be represented as a binary matrix indicating whether an individual possesses a particular disease. Clustering algorithms can help us to identify recurrent multimorbid disease groupings.

natively, Planell-Morell et al. (2020) used a Markov graph clustering approach to understand the sequence of conditions acquired by individuals, whilst recent deep learning approaches have been developed to learn representations of heterogeneous, non-uniformly sampled clinical data (Qian et al., 2020).

2. Contribution

In this paper, we address the related problem where access to EHRs is limited or health survey data is being used. In these situations, full temporal information is not available and only a *cross-sectional* snapshot of the study population is recorded. This would provide information about the conditions an individual possesses *at the time of* the survey but no detailed health history. The cross-sectional data scenario is illustrated by our main exemplar, the Golestan study (Odland et al., 2021). Existing computational phenotyping approaches are not applicable in this situation.

We assume cross-sectional multimorbidity data is available in the form of a binary individual-condition matrix (Figure 1). Further we will also assume that age information is available for each individual which we shall leverage to reconstruct temporal patterns in multimorbidity acquisition in the absence of explicit diagnosis times.

In this work, we construct a time-evolving Bayesian feature allocation model which explicitly incorporates age-dependence in the accumulation of multimorbidities. Our contributions are to 1) Develop the first bespoke probabilistic generative model that allows one to make inferences about the temporal dependencies driving patterns of multimorbidity without having to follow any specific individual through time. 2) Demonstrate an efficient inference algorithm which is exact (there is no model approximation during inference). 3) Show how our model can provide a clear representation of clinically meaningful multimorbidity patterns comprising explicit disease profiles and prevalence trajectories thus facilitating interpretable analysis in both simulated and real-world examples.

3. Wright–Fisher Multimorbidity Trajectory Model

In the following, we describe the construction of our model, which we call the *Wright–Fisher Multimorbidity Trajectory Model* (WF-MTM).

3.1. Generative model

We assume the existence of K latent time-varying features, where $X_k(t)$ denotes the probability that, at time t, the latent feature k would be active. The prior over each feature trajectory is given by a Wright–Fisher diffusion with mutation parameters $\alpha\beta/K$ and β (see e.g. Durrett, 2008, Ch. 7):

$$X_k(t) \sim \mathrm{WF}\left(\frac{\alpha\beta}{K}, \beta\right),$$

where α and β control the prevalence of the latent features.

We associate each latent feature with an intensity $\phi_k \sim \mathrm{Gamma}(\gamma, 1)$ and a *morbidity*

probability profile, ρ_k, which tells us the probability of a morbidity d occurring for feature k. Its prior is:

$$\rho_{kd} \sim \text{Beta}(\eta_0, \eta_1), \qquad d = 1, \ldots, D.$$

Thus each latent feature can be considered to be associated with a collection of comorbidities. The hyperparameters η_0, η_1 are taken to be identical across all features, implying that the distribution of each feature being associated with each morbidity is the same *a priori*. In order to preserve model identifiability, we allow latent feature prevalence to evolve with time but not their morbidity profiles.

Given the latent feature trajectories, a binary feature indicator $\mathbf{Z}$ assigns latent features to individuals at each time point. From this, we specify $\boldsymbol{\theta}_{it} = (\theta_{itk})$ which is a probability distribution over the K latent features for each individual i at time t:

$$Z_{itk}|X_k(t) \sim \text{Bernoulli}(X_k(t)),$$
$$\boldsymbol{\theta}_{it}|\mathbf{Z}_{it}, \boldsymbol{\phi} \sim \text{Dirichlet}(\mathbf{Z}_{it} \circ \boldsymbol{\phi}),$$

where $\mathbf{Z}_{it} = (Z_{itk})$, $\boldsymbol{\phi} = (\phi_k)$, and $\mathbf{Z}_{it} \circ \boldsymbol{\phi}$ refers to their Hadamard product. This construction allows the model to distinguish between feature prevalence and its contribution to morbidity within an individual. Thus we can successfully account for the possibility of a morbidity profile which is rare in the population but which nonetheless accounts for a substantial fraction of the conditions within those individuals associated with that profile (Williamson et al., 2010).

Finally, we use $W_{itd} \in \{0, 1\}$ to indicate whether individual i at time t has morbidity d. The morbidity is assumed to derive from a latent feature and we use $A_{itd} \in \{1, \ldots, K\}$ as an indicator:

$$A_{itd}|\boldsymbol{\theta}_{it} \sim \text{Categorical}(\boldsymbol{\theta}_{it}),$$
$$W_{itd}|A_{itd}, \rho_{A_{itd}} \sim \text{Bernoulli}(\rho_{A_{itd}d}).$$

A graphical illustration of the model is given in Figure 10.

3.2. Inference

Posterior inference is based on Markov Chain Monte Carlo (MCMC) with Gibbs-type updates where the algorithm cyclically samples a single variable from its conditional distribution given the others.[1] To improve mixing time, we integrate out θ and ρ and sample only A, $\boldsymbol{\phi}$, Z, and X following Perrone et al. (2017). The allocations A can be sampled with a Gibbs step and the parameters $\boldsymbol{\phi}$ can be straightforwardly updated with a random-walk Metropolis step (see Appendix). However, updating Z and X is more complicated, as we now describe.

Sampling Z. The posterior distribution of individual latent feature indicators $\mathbf{Z}_{it}$ depends on the posterior distribution of the number of feature assignments across individual morbidities, $\mathbf{n}_{it} = (n_{itk})$. Suppose $\boldsymbol{\phi}^* = (\phi_s)$, $\mathbf{n}^*_{it} = (n_{its})$ are S-dimensional sub-vectors with subscript s, the sth largest element of the set $S_{it} = \{k : Z_{itk} = 1\}$ of the cardinality S. Then

$$P(\mathbf{Z}_{it}|\mathbf{n}_{it}, \mathbf{X}(t), \boldsymbol{\phi}) \propto \binom{D}{\mathbf{n}^*_{it}} \frac{B(\boldsymbol{\phi}^* + \mathbf{n}^*_{it})}{B(\boldsymbol{\phi}^*)} \times \prod_{k=1}^{K} X_k(t)^{Z_{itk}} (1 - X_k(t))^{1 - Z_{itk}},$$

where $B(\cdot)$ is the beta function.

We use a Hamming Ball sampler for efficient sampling of $\mathbf{Z}_{it}$ (Titsias and Yau, 2017); this employs an auxiliary variable $\mathbf{U}_{it}$ that allows iterative sampling from slices of the state space of $\mathbf{Z}_{it}$ and avoids exhaustive enumerations over the entire state space of $\mathbf{Z}_{it}$ with exponential complexity. The Hamming Ball sampler can be more effective than the standard Gibbs sampler in inferring a correlated posterior distribution, e.g. where latent groups of morbidities co-occur within

1. We provide implementation of our inference at https://github.com/thysics/WF-MTM.

Reference	Survey	Temporal Data	Temporal Model	Feature Allocation	Uncertainty Quantification
Clustering algorithms	Cross-sectional	✗	✗	✗	✗
Ruiz et al. (2014)	Cross-sectional	✗	✗	✓	✓
Hassaine et al. (2020)	Longitudinal	✓	✓	✗	✗
Qian et al. (2020)	Longitudinal	✓	✓	n/a	✓
WF-MTM	Cross-sectional	✗	✓	✓	✓

Table 1: **Existing approaches for multimorbidity analysis.**

an individual, since it has the ability to perform joint updates and move between different posterior modes. The algorithm consists of two steps:

$$\mathbf{U}_{it} \leftarrow P(\mathbf{U}_{it}|\mathbf{Z}_{it}), \tag{1}$$
$$\mathbf{Z}_{it} \leftarrow P(\mathbf{Z}_{it}|\mathbf{U}_{it}, \Theta), \tag{2}$$

where $\Theta = \{\mathbf{n}_{it}, X_t, \boldsymbol{\phi}\}$.

The conditional distribution of $\mathbf{U}_{it}$ is

$$P(\mathbf{U}_{it}|\mathbf{Z}_{it}) = \frac{1}{Z_m}\mathbb{I}(\mathbf{U}_{it} \in \mathbb{H}_m(\mathbf{Z}_{it}))$$

where $\mathbb{H}_m(\mathbf{Z}_{it}) = \{\mathbf{U}_{it} : \sum_{k=1}^{K}\mathbb{I}(u_{itk} \neq z_{itk}) \leq m\}$ and Z_m is the cardinality of $\mathbb{H}_m(\mathbf{Z}_{it})$. The posterior conditional $P(\mathbf{Z}_{it}|\mathbf{U}_{it}, \Theta)$ simplifies to $P(\mathbf{Z}_{it}|\mathbf{U}_{it}, \Theta) \propto P(\mathbf{Z}_{it}, \mathbf{U}_{it}, \Theta)\mathbb{I}(d(\mathbf{U}_{it}, \mathbf{Z}_{it}) \leq m)$ where $d(\cdot)$ denotes Hamming distance and m is the user-defined Hamming radius, i.e. the number of elements in $\mathbf{U}_{it}$ we allow to differ from $\mathbf{Z}_{it}$.

Sampling X. The posterior distribution of feature probabilities X across time is proportional to the transition probability of the WF diffusion and is intractable. We use Particle Gibbs sampling (Andrieu et al., 2010) to approximate this posterior distribution. More precisely, we follow Perrone et al. (2017, Algorithm 1) except we replace their discretisation of the WF diffusion with an *exact* WF simulation method to eliminate all discretisation error. The WF diffusion can be simulated exactly by exploiting a probabilistic representation of the transition function's eigenfunction expansion (Jenkins and Spanò, 2017).

By incorporating this exact simulation method, we gain considerable computational advantages compared to previous approaches. While the exact method has time complexity $O(T)$ such that T refers to the number of age cohorts, a discretisation method has higher complexity of $O(GT)$ where G is the number of grid points. In practice this translates into an order of magnitude in computational gains.

4. Related Work

Despite its wide use in different fields such as natural language processing (Williamson et al., 2010; Perrone et al., 2017), public health (Ruiz et al., 2012), and disease phenotyping (Ni et al., 2020), Bayesian feature allocation models seem to be applied rarely in multimorbidity analysis. An exception is the work of Ruiz et al. (2014). This method is closest to ours since it is not only applicable to binary data but also can be used to obtain an interpretable representation of groups of morbidities. However, this model does not incorporate age information and therefore is not suited for identifying patterns of age-dependent co-occurring diseases.

Previous studies have also used generic clustering methods to partition patients into mutually exclusive (latent) sets. These include the use of K-means (KM) (Violán et al., 2019), Hierarchical Clustering Anal-

ysis (HCA) (Roso-Llorach et al., 2018), and Latent Class Analysis (LCA) (Larsen et al., 2017; Hall et al., 2018; Zhu et al., 2020). However, with such approaches, the time-evolving behaviour of clusters can only be examined by partitioning the study population into age groups, clustering each age group individually, and then integrating outputs post-hoc.

5. Experiments

We demonstrate the utility of our model on both simulated and real-world data. For comparison, we use a set of benchmark models which are widely used in the analysis of multimorbidity: KM, HCA, LCA, and an Indian Buffet Process-based feature allocation model (IBP) (Ruiz et al., 2014). The evaluation is based on whether each model can identify clinically meaningful latent morbidity profiles and whether these profiles exhibit time-dependence in their prevalence. In curated examples, we compare the outcome of each model with respect to a (curated) ground truth; otherwise, we use other data-driven metrics for comparison. To better understand how these models would perform in real-world applications, we considered a range of examples with increasing complexity.

To compare methods we must account for the fact that no model except for ours explicitly models the temporal dependence of multimorbidity patterns. We therefore conduct additional 'post-analysis curations' on comparator methods in order to obtain a proxy for the trajectory of a feature and its morbidity profile. We construct trajectories retrospectively as follows: The morbidity profile associated with a latent feature k is constructed by looking only at those individuals to whom feature k—and no other features—are associated. The empirical means of morbidities within these observations are then used as the corresponding morbidity profile. For clustering methods, a temporal trajectory is constructed by tracking the empirical size of a cluster at each time where the cluster itself is identified without knowing a patient's age. We emphasise that our method does not require such ad-hoc post-analysis curation and *automatically* provides a clear probabilistic representation of time-dependent latent morbidity profiles.

Curated example: Non-overlapping morbidity profiles. We considered a complex simulated example which mimics a real cross-sectional study. In this example, each patient accumulates a series of symptoms (out of 15 different morbidities) from one or more latent multimorbidity features throughout their lifetime (10% of the population develop morbidities from two different features). Latent features have time-varying prevalence. For instance, a collection of co-morbidities (numbered 5–9) appears in early age groups while a separate feature associated with morbidities numbered 10–14 emerges from middle age. The third profile, associated with morbidities 0–4, is present across all ages (Figure 7). To better reflect the real-world data, 20% of patients are assumed to be morbidity-free.

We further allow the morbidity profile within each feature to be time-dependent, $\rho_{kd} = \rho_{kd}(t)$. This deviates from our generative process which assumes a time-invariant profile where only the feature prevalences $X_k(t)$ evolve over time. Inference under this example therefore also serves as a robustness check for model misspecification.

Results are shown in Figure 2(a). Our method, WF-MTM, as well as IBP successfully recover morbidity profiles. Where they are time-dependent, the inferred profiles are close to an average of underlying profiles across time. However other, clustering-based, methods fail to recover un-

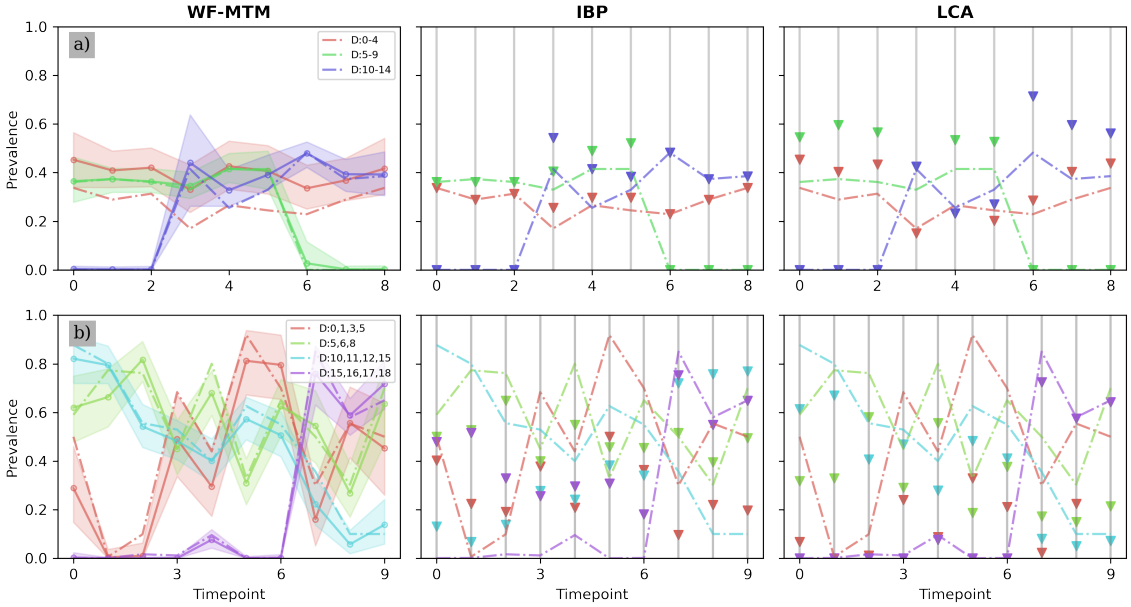

Figure 2: **Analysis of simulated cross-sectional data sets.** Two types of simulated examples: a) non-overlapping and b) overlapping morbidity profiles. Empirical proportion of patients in each feature (dashed line). WF-MTM: posterior credible interval and its mean (shaded area, bold line) for feature trajectories. IBP and LCA: a proportion of each feature at each age (inverted triangle); note that the models themselves are not time-dependent.

derlying profiles since they all cluster *individuals* rather than *morbidities*. Their inaccurate reconstructions of morbidity profiles lead to the overestimation of feature temporal prevalence. Figure 2(a) shows that unlike feature-allocation models, LCA overestimates the prevalence of features across time (similarly for KM and HCA). While both WF-MTM and IBP reasonably recover temporal trajectories, it is important to remark that only our method can make inferences for such trajectories directly without the need for additional analysis.

Overlapping morbidity profiles. We considered another simulated example where we simulate 913 patients from our generative process across 10 time (i.e. age) points. Here patients are susceptible to four underlying time-varying features whose profiles involve $D = 20$ morbidities. Profiles were constructed in a way that each pair of profiles share one common morbidity. In this more realistic scenario the majority of patients, 68% of the population, possess morbidities from multiple features with overlapping morbidity profiles.

Posterior samples from the MCMC algorithm successfully recapitulate the underlying (true) feature trajectories and morbidity profiles. Figure 2(b) shows that the credible intervals for trajectories generally capture the corresponding truth. However, the other baseline methods, including IBP, reconstruct morbidity profiles and temporal prevalences which significantly deviate from the underlying truth. For instance, LCA was only able

Data	Model	Pearson	Spearman
Semi-curated data	WF-MTM	**[0.360, 0.394]**	**[0.337, 0.378]**
	IBP [2]	0.190	0.167
	LCA	[0.310, 0.373]	[0.305, 0.371]
	HCA	[0.035, 0.331]	[0.043, 0.332]
	KM	[0.276, 0.354]	[0.272, 0.357]
Real-world data	WF-MTM	**[0.362, 0.389]**	**[0.343, 0.372]**
	LCA	[0.268, 0.340]	[0.258, 0.330]
	HCA	[0.061, 0.207]	[0.050, 0.189]
	KM	[0.145, 0.296]	[0.137, 0.288]

Table 2: **Performance for all baselines.** Performance intervals are based on either the 95% posterior credible interval (WF-MTM) or the corresponding bootstrap confidence interval (others). Best performance is highlighted in bold. See Appendix for definitions of correlation metrics and bootstrap confidence interval.

to reconstruct the trajectory of the feature shown in purple while its estimates for others continue to under-estimate the ground truth.

Golestan study. Odland et al. (2021) sampled individuals aged 36–81 years in the Golestan province in Iran. The data was collected from a cross-sectional cohort study conducted between 2006 and 2010 where each individual is recorded only once. We filtered the data so that every patient has at least two morbidities. This resulted in $N = 13,953$ patients with 37 different age-groupings from 39–75 years. On average, each patient possesses 2.6 of the $D = 19$ conditions and there are $1,534$ unique sets of co-morbidity patterns within the data, whose size ranges from two to nine conditions. We applied our method to infer age-dependent multimorbidity profiles from the cross-sectional data.

Semi-curated dataset. Before analysing the full data, we considered a subset which we refer to as the semi-curated dataset, consisting of patients exhibiting few distinct multimorbidity patterns. Specifically, we sample a subset of individuals each of whom possesses either hypertension, thyroid disease, or stroke. This experiment allows us to compare the performance of our model against the current state-of-the-art feature allocation-based model (IBP).[3]

Furthermore, unlike the full dataset where there is no information about the ground truth, the curated set allows us to seed some anticipated multimorbidity patterns, which makes it possible to evaluate the outcome of each model. A reasonable reconstruction of morbidity profile and feature trajectory is expected to be consistent with existing clinical knowledge. For instance, in Iran, the stroke incidence rate is higher among the older population and the condition tends to co-occur with hypertension (Fallahzadeh et al., 2022). Other studies suggest that the majority of thyroid problems in developing countries occur in young adults (e.g. 32–51 years) (Tahir et al., 2020; Tsegaye and Ergete, 2003). It is also well-known that there is a strong association between hypertension and thyroid problems (Berta et al., 2019).

In our comparison, WF-MTM, IBP and LCA are able to summarize dominant multimorbidity patterns in the form of three distinct latent features, each of which captures a set of co-occurring diseases coupled with stroke, thyroid disease, and hypertension (Figure 3). However, it was only possible to do so for IBP and LCA through manual stratification of patients into distinct age groups. K-means and hierarchical clustering did not reproduce these patterns. A morbidity profile associated with thyroid disease

2. We could not obtain the outcome based on the posterior credible interval for IBP since this requires substantial modifications of its existing code base.

3. We could not apply IBP to the full dataset since we found that its existing code base is scalable only up to a couple of thousand observations.

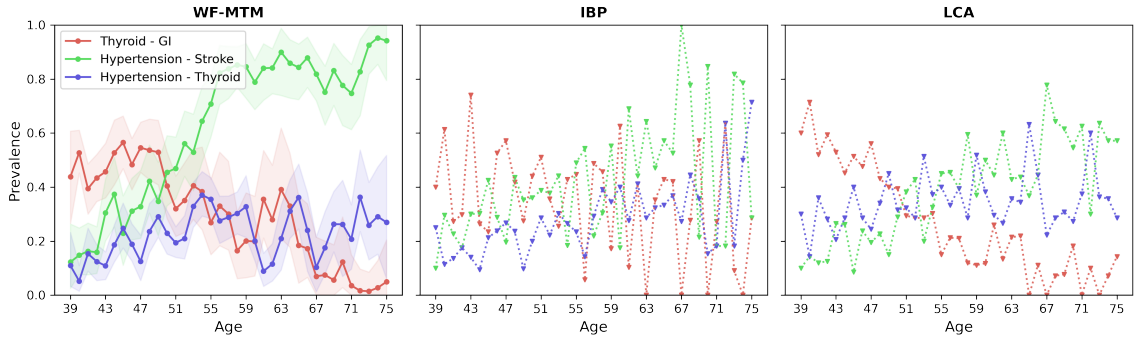

Figure 3: **Analysis of semi-curated dataset.** The posterior credible interval and its mean (shaded area, bold line) for each feature inferred by WF-MTM. A proportion of each feature at each age, constructed from the baseline models (inverted triangle, dotted line). Features are named after their two leading conditions.

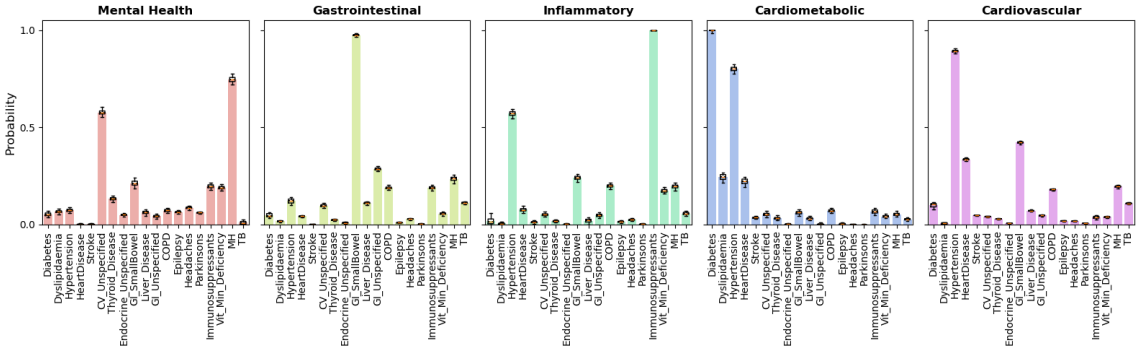

Figure 4: **Multimorbidity disease profiles.** The posterior means of disease occurrence probabilities in each feature. Boxes cover the interquartile interval of the estimates with whiskers being the 95% credible interval.

(or stroke) is strongly correlated with hypertension, agreeing with well-known disease associations (Figure 9). Our model, however, stood out by its ability to recapitulate *age-linked* dependencies. As expected, Table 2 shows that the features inferred by WF-MTM possess temporal dependencies that better reflect changes in the occurrences of its member diseases across time than other baseline models. Furthermore, the posterior feature trajectories are easily interpretable thanks to their stable and smooth shape, different from the ones obtained by IBP, which are both rough and unstable and therefore harder to interpret (Figure 3). The difference is mainly due to explicit modelling of time-dependence in WF-MTM allowing the sharing of statistical strength across age-groupings, which encourages the emergence of multimorbidity patterns with clearer temporal dependency.

Full real-world data. We next considered the full real-world data. Incorporating age

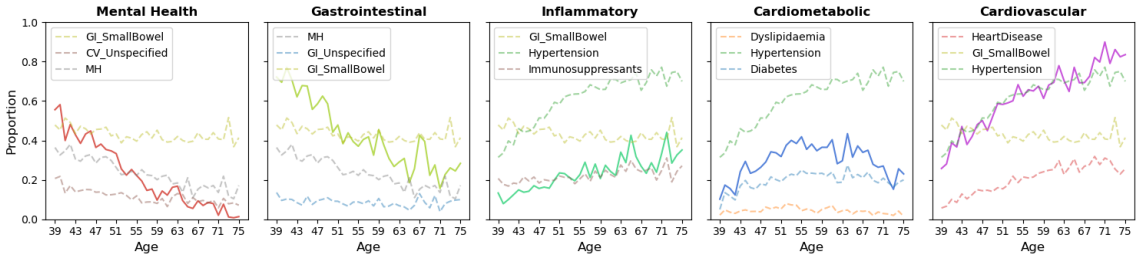

Figure 5: **Temporal prevalence of single conditions.** Each panel includes empirical trajectories of single conditions (dashed lines) selected from the three most common diseases within that morbidity profile. Inferred feature trajectory is in bold.

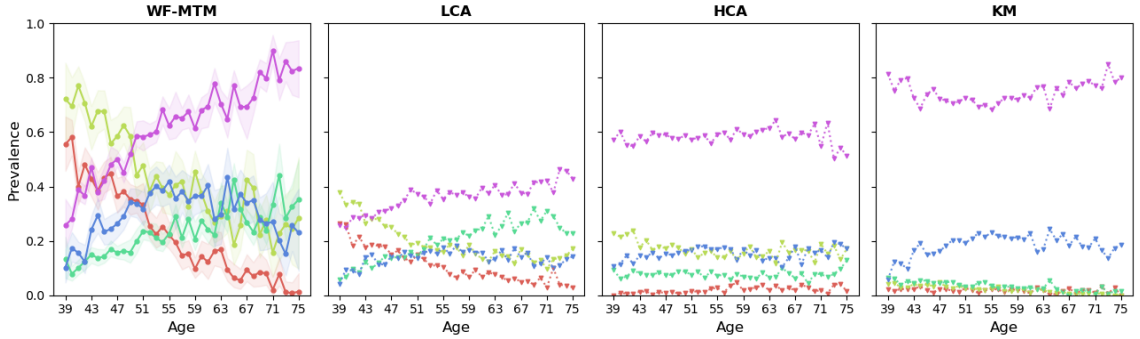

Figure 6: **Reconstruction of feature prevalence trajectory.** Each coloured dotted line represents the age-dependent prevalence trajectory of a feature. Trajectories in benchmark models are constructed based on the empirical proportion of each feature at each age. Features from various models are matched with each other (by colour) in terms of their closeness with respect to the Wasserstein metric.

allows WF-MTM to identify time-dependent emergence of morbidity profiles such as mental health (MH), gastrointestinal (GI), inflammatory (INF), cardiometabolic (CM) and cardiovascular (CV) profiles (we name each profile after their leading constituent disease). The patterns of co-occurrence among diseases are consistent with well-known disease-associations. Figure 4 shows for example that feature CM captures that hypertension is coupled with heart disease and diabetes, while various gastrointestinal conditions are grouped together in GI. Furthermore, patient-level covariate data—not used in the analysis—further support the validity of our findings; for instance, relative to the population average, the 'inflammatory' feature is associated with increased numbers with arthritis (289% higher) whereas both CV and CM are associated with increased systolic blood pressure (20.6% and 22% higher, respectively) in comparison with GI (the feature associated with the lowest blood pressure).

The reconstructed temporal prevalences from WF-MTM are also consistent with the

temporal prevalence of its leading conditions in the data (Figure 5). For instance, the 'mental health' profile, which shows a prevalence that decreases with age, has two leading conditions including 'mental health disorder' and 'unspecified cardiovascular disease', both of whose occurrences decrease over time. The cardiovascular profile, on the other hand, has an upward-sloping trajectory, following that of its leading conditions such as hypertension and heart disease.

Our model demonstrates superior performance to alternative baseline models in capturing the age-dependency of disease accumulation. Figure 6 shows that the implied prevalence of each feature obtained from KM and HCA have little dependence on age while the prevalence trajectories from our model can capture strong dependency across age. This is supported by a quantitative assessment: Table 2 shows that our model performed the best in each metric.

6. Conclusion

We introduced the Wright–Fisher Multimorbidity Trajectory Model (WF-MTM), a novel probabilistic generative model for multimorbidity analysis. Our model frames multimorbidity analysis as a latent feature allocation problem which assumes individual disease susceptibility is driven by multiple latent factors. We introduce temporal dependence on these latent features, allowing for age-varying multimorbidity trajectories. We achieve state-of-the-art performance on both synthetic and real-world data examples. Using cross-sectional health surveys of large cohorts with binary records of disease conditions as well as patient ages, we are able to identify age-dependent properties of latent features driving the acquisition of multimorbidity.

The model we introduce is flexible and easily extended to a nonparametric counterpart in which the number of clusters, K, can be regarded as unknown and unbounded. We found our model to be reasonably robust to misspecification of K (Figure 8) and to other types of model misspecification as discussed above, though we leave detailed analysis of the inference of K to future work. Another extension includes a comparison between reconstructed groupings of morbidities from the model and known biomedical classifications, e.g., morbidity groupings that are known to co-exist. Other important extensions include allowing for time-dependence of morbidity profiles within features, and the incorporation of patient covariate data, beyond just age, directly into the model.

Acknowledgments

We would like to thank Professor Justine Davies (University of Birmingham) for providing us with invaluable, curated data for the Golestan Study and Dr. Francisco J. R. Ruiz for providing us with the implementation code for his work (Ruiz et al., 2014). Permission to utilise the Golestan Study was granted on 2022–03–03 under GEMShare Application #245. PJ is supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. WK is supported by the CDT in Mathematics and Statistics at the University of Warwick and the Enrichment Scheme from the Alan Turing Institute (EPSRC Grant Ref: EP/N510129/1). CY is supported by an EPSRC Turing AI Acceleration Fellowship (Grant Ref: EP/V023233/1). CY acknowledges support from the NIHR Programme for Artificial Intelligence for Multiple Long-Term Conditions (Grant Ref: NIHR202632).

References

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain

Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Eszter Berta, Inez Lengyel, Sándor Halmi, Miklós Zrínyi, Annamária Erdei, Mariann Harangi, Dénes Páll, Endre V Nagy, and Miklós Bodor. Hypertension in thyroid disorders. *Frontiers in endocrinology*, 10: 482, 2019.

Richard Durrett. *Probability models for DNA sequence evolution*, volume 2. Springer, 2008.

Aida Fallahzadeh, Zahra Esfahani, Ali Sheikhy, Mohammad Keykhaei, Sahar Saeedi Moghaddam, Yeganeh Sharifnejad Tehrani, Negar Rezaei, Erfan Ghasemi, Sina Azadnajafabad, Esmaeil Mohammadi, et al. National and subnational burden of stroke in iran from 1990 to 2019. *Annals of clinical and translational neurology*, 9(5):669–683, 2022.

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

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

Paul A Jenkins and Dario Spanò. Exact simulation of the Wright–Fisher diffusion. *The Annals of Applied Probability*, 27(3):1478–1509, 2017.

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.

Drew A Linzer and Jeffrey B Lewis. polca: An r package for polytomous variable latent class analysis. *Journal of statistical software*, 42:1–29, 2011.

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

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

Valerio Perrone, Paul A Jenkins, Dario Spanò, and Yee Whye Teh. Poisson random fields for dynamic feature models. *Journal of Machine Learning Research*, 18, 2017.

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

Zhaozhi Qian, Ahmed Alaa, Alexis Bellot, Mihaela Schaar, and Jem Rashbass. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3295–3305. PMLR, 2020.

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

Francisco Ruiz, Isabel Valera, Carlos Blanco, and Fernando Perez-Cruz. Bayesian nonparametric modeling of suicide attempts. *Advances in neural information processing systems*, 25, 2012.

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

Noor Thair Tahir, Hadeel Delman Najim, Aufaira Shaker Nsaif, et al. Prevalence of overt and subclinical thyroid dysfunction among iraqi population in baghdad city. *Iraqi Journal of Community Medicine*, 33 (1):20, 2020.

Michalis K Titsias and Christopher Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, 112 (520):1598–1611, 2017.

B Tsegaye and W Ergete. Histopathologic pattern of thyroid disease. *East African medical journal*, 80(10):525–528, 2003.

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.

Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

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

Appendix A. Posterior inference

Sampling A. The posterior distribution of the feature assignment for morbidity d in individual i at time t, A_{itd}, is proportional to the conditional distribution of $\mathbf{A}_{-itd} = (A_{i't'd'})_{(i',t',d')\neq(i,t,d)}$, i.e. of the cluster assignments for all the other morbidities except for A_{itd}, times the prior distribution of the corresponding morbidities status W_{itd}.

$$
\begin{aligned}
&P(A_{itd}=k|\mathbf{A}_{-itd},\mathbf{W},\mathbf{Z}^{\mathbf{it}}_{-\mathbf{k}},Z_{itk}=0,\boldsymbol{\phi})=0,\\
&P(A_{itd}=k|\mathbf{A}_{-itd},\mathbf{W}_{-itd},W_{itd}=1,\mathbf{Z}^{\mathbf{it}}_{-\mathbf{k}},Z_{itk}=1,\boldsymbol{\phi})\\
&\propto \frac{(\phi_k+n_{itk}-\mathbb{I}(A_{itd}=k))(\eta_0+n^{kd}_{v=1}-\mathbb{I}(W_{itd}=1,A_{itd}=k))}{\eta_0+\eta_1+\sum_{j=0}^{1}(n^{kd}_{v=j}-\mathbb{I}(W_{itd}=j,A_{itd}=k))},\\
&P(A_{itd}=k|\mathbf{A}_{-itd},\mathbf{W}_{-itd},W_{itd}=0,\mathbf{Z}^{\mathbf{it}}_{-\mathbf{k}},Z_{itk}=1,\boldsymbol{\phi})\\
&\propto \frac{(\phi_k+n_{itk}-\mathbb{I}(A_{itd}=k))(\eta_1+n^{kd}_{v=0}-\mathbb{I}(W_{itd}=0,A_{itd}=k))}{\eta_0+\eta_1+\sum_{j=0}^{1}(n^{kd}_{v=j}-\mathbb{I}(W_{itd}=0,A_{itd}=k))}.
\end{aligned}
$$

Here the number of morbidity 'locations' assigned to feature k and resulting in morbidities status v (either 0 or 1) is denoted by

$$n^{kd}_{v}=\sum_{t=1}^{T}\sum_{i=1}^{N_t}\mathbb{I}(A_{itd}=k,W_{itd}=v),$$

and $\mathbf{Z}^{it}_{-k} = (Z_{itj})_{j \neq k}$. The above conditional distribution can be computed exactly by enumeration across all $k = 1, \ldots, K$; this update is a Gibbs step.

Sampling $\boldsymbol{\phi}$. The posterior distribution of $\boldsymbol{\phi}$ depends on the distribution of the feature assignments across the population and on the prior distribution of $\boldsymbol{\phi}$. Let $\boldsymbol{\phi}^* = (\phi_s), \mathbf{n}^*_{it} = (n_{its})$ be S-dimensional vectors, where S denotes the cardinality of the set $S_{it} = \{k : Z_{itk} = 1\}$. To simplify the notation, we use the same notation S for every observation although S varies across the population. Then

$$P(\boldsymbol{\phi}|\gamma, \mathbf{A}, \mathbf{Z}) \propto \prod_{k=1}^{K} \frac{\phi_k^{\gamma-1} e^{-\phi_k}}{\Gamma(\gamma)} \times \prod_{t=1}^{T} \prod_{i=1}^{N_t} \binom{D}{\mathbf{n}^*_{it}} \frac{B(\boldsymbol{\phi}^* + \mathbf{n}^*_{it})}{B(\boldsymbol{\phi}^*)}$$

where $B(\cdot)$ is the multivariate Beta function. We obtain samples from this distribution based on a random-walk Metropolis–Hastings algorithm with a normally distributed proposal whose variance is set to 0.1 throughout.

Appendix B. Implementation Details

Clustering methods. The benchmark clustering algorithms include K-means (KM), Hierarchical Clustering Analysis (HCA) and Latent Class Analysis (LCA). We apply KM on the space of principal components of the data. We use the Jaccard metric for HCA. The implementation of LCA follows Linzer and Lewis (2011).

IBP. The benchmark feature-allocation model is developed in Ruiz et al. (2012, 2014). Inference is based on a Gibbs-sampler (Ruiz et al., 2012) and the implementation code was provided by the author. Since this model requires data containing subjects without any morbidities, we included those in both curated and real-world data. Latent feature memberships of subjects without any morbidity is manually set to be the zero vector, following Ruiz et al. (2014). Although this method can infer the *number* of latent features, we found that the inferred number of latent features significantly deviated from the ground truth in both curated and semi-curated datasets. Therefore, we fixed the number of latent features to be the true value throughout.

WF-MTM. Throughout the paper, we use the hyper-parameter set-up: $\eta_0 = \eta_1 = 0.2$, $\gamma = 1.5$, $\alpha = \frac{1}{K}$, $\beta = 1$, where K is the number of latent features. The diffusion time unit, $dt_j = t_j - t_{j-1}$, $\forall j$ is 0.5 (Inference under the model) or 0.05 (The rest). The posterior analysis is conducted based on the Monte Carlo samples obtained between 4000–5000th iteration (Semi-curated example) and 2000–3000th iteration (Real-world example).

Unlike IBP, our model is applicable both with or without morbidity-free subjects; for instance, in the curated example, our model infers latent feature membership of those without morbidity, and this leads to the emergence of a "healthy" feature, i.e. the one exclusively assigned to subjects without any morbidity. In the main analysis, however, we manually assigned zero latent features to subjects without morbidity for comparison with IBP.

Performance metric error measurements To quantify the uncertainty of performance evaluation measurements, we used a 95% bootstrap confidence interval for baseline methods including KM, HCA and LCA whereas the 95% posterior credible interval was used for our model.

Bootstrap confidence intervals were constructed by the following steps: after creat-

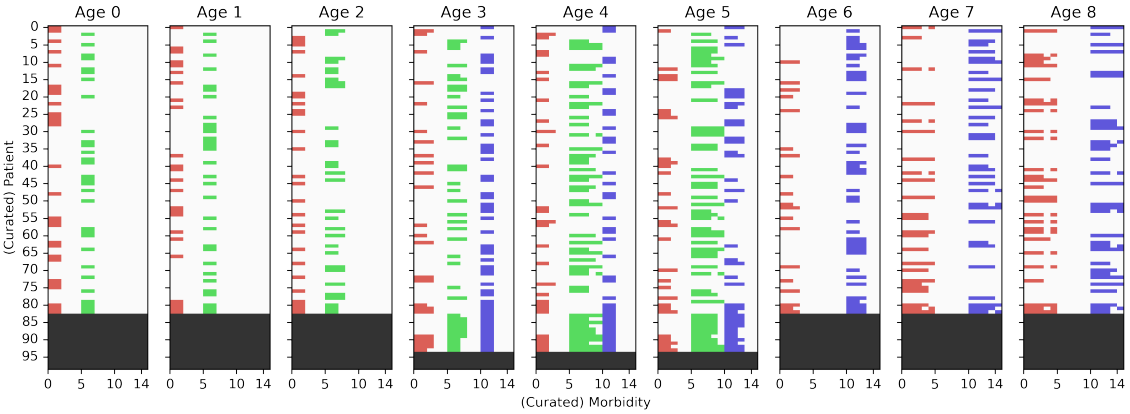

Figure 7: **Simulated cross-sectional data set.** Each row represents a simulated patient with a set of binary morbidity indicators across 15 diseases where zero values are shown in white and those in color correspond to their comorbidity membership. As in a real cross-sectional cohort, there are variable numbers of individuals in each of the nine age groupings.

ing replicates of the dataset with sampling with replacement, each baseline method is implemented on each replicate where the cluster trajectories are then obtained via post-analysis curation. Based on these trajectories, we compute the correlation metrics. The 95% bootstrap confidence interval is comprised of the values between 2.5% quantile and 97.5% quantile correlation values across the entire set of replicates.

Appendix C. Sensitivity Analysis

Time-varying disease profile. As described in the main article, we curated an example where each patient acquires disease sequentially based on their feature membership. Figure 7 shows that there are three distinct groups of patients and their feature membership is indicated in corresponding colours. According to their morbidity profile, each patient accumulates morbidities in chronological order; for instance, a green group is likely to acquire the 5th morbidity at a young age and then the 9th one sequentially.

Correlation metric. Suppose K is the number of latent co-morbidity profiles (i.e. features) and D is the number of morbidities. T denotes the number of age cohorts. The correlation performance metrics are defined as

$$\frac{1}{K}\sum_{k=1}^{K} p_k \frac{\sum_{d=1}^{D}\rho_{kd}\mathrm{Corr}(\mathbf{X}_k,\mathbf{E}_d)}{\sum_{d=1}^{D}\rho_{kd}} \Big/ \sum_{k=1}^{K} p_k, \qquad (3)$$

where $\mathbf{X}_k = (X_{kt})_t$ is estimated feature trajectory. $E_d = (E_{dt})_t$ refers to the empirical proportion of individuals with the dth morbidity across all age cohorts, i.e. $E_{dt} = \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{I}(W_{idt} = 1)$ such that W_{idt} is the dth morbidity indicator of the ith individual in the tth age cohort. ρ_{kd} is the estimated morbidity acquisition probability for the dth morbidity in the kth feature. p_k is the mean of the estimated cluster prevalence across time, i.e. $p_k = \frac{1}{T}\sum_{t=1}^{T} X_{kt}$.

We choose these correlation metrics mainly because of their familiarity and sim-

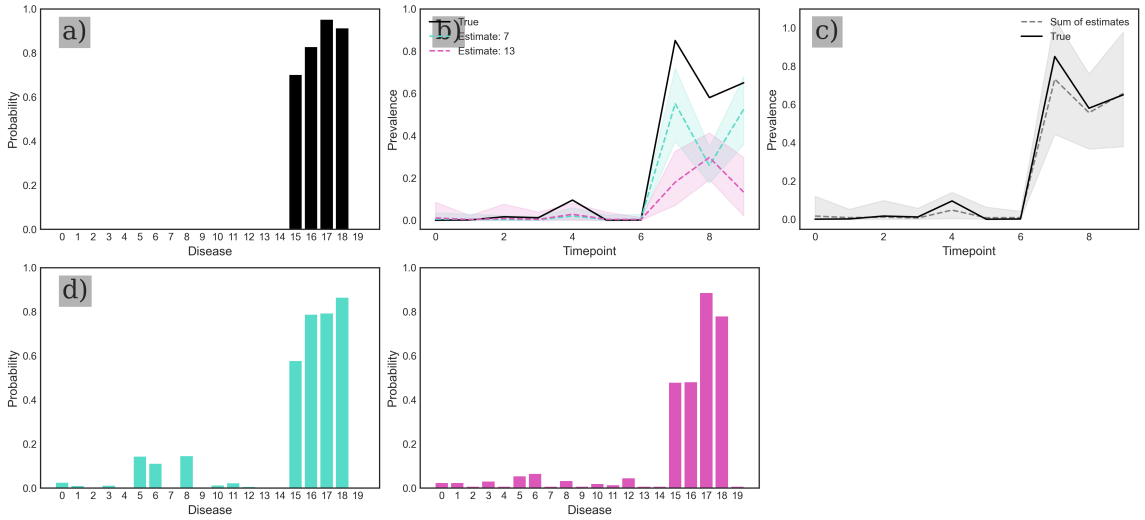

Figure 8: **Inference under model mis-specification.** **(a)** True morbidity profile. **(b)** True feature trajectory in black while relevant trajectory estimates are in colour. **(c)** True trajectory in bold and the sum of posterior trajectories is in dashed line with 95% credible interval in shaded area. **(d)** Reconstructed morbidity profiles combined to obtain the trajectory in (c).

plicity. The correlation between estimated temporal co-morbidity trajectory and empirical prevalence of each single condition serves as a good first-order approximation to their relationship and therefore can be used to compare different models in terms of their ability to recapitulate age-linked dependencies of co-morbidities. Note that this performance metric is unique to our analysis since existing works did not consider age information in finding correlations between morbidities.

Over-specification of K. We conducted an experiment where we deliberately over-specify the number of latent features K. Since there is no information about "true" latent multimorbidity in real-world data, it is important that inferred features capture a similar set of morbidities even when K is mis-specified. We fit our model on the simulated data in the previous section where we mis-specify $K = 15$, more than three times larger than the true value ($K = 4$). Apart from this, we use the same set of hyperparameters as before. Our goal is to examine (i) whether the reconstruction of true morbidity profiles is achievable, and (ii) if it is possible to recover underlying trajectories.

When K is mis-specified in this way, we recover the true comorbidity profiles of all features but in the form of a posterior ρ of a set of 'duplicate' features where their co-morbidity profiles are either equal or very similar to each other. Furthermore, we can even reconstruct the "true" prevalence trajectory of each feature by summing up the trajectories of the set of corresponding duplicate features. Figure 8 shows a true feature along with a set of its estimates. These estimates have comorbidity profiles almost equal to their 'target' profile while the sum of their estimated trajectories encapsulate the true trajectory in their 95% credible interval.

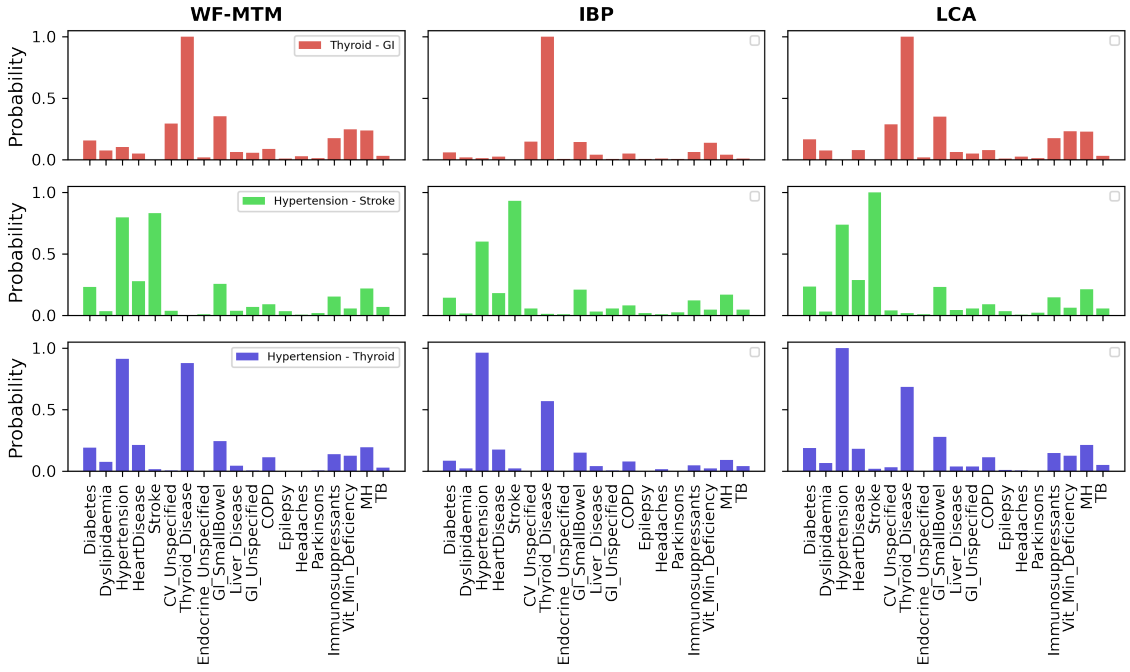

Figure 9: **Analysis of semi-curated dataset: morbidity profiles.** Each column shows reconstructed morbidity profiles from each model in the semi-curated dataset.

This demonstrates that the model is reasonably robust to the mis-specification of K.

Appendix D. Additional Figures

Figure 9 illustrates reconstructed morbidity profiles from each model in the semi-curated dataset. These profiles seem to be consistent with well-known disease-associations, e.g. hypertension–thyroid problems and hypertension–stroke.

Figure 10 shows a graphical illustration of our model. We first assume that there exist a fixed number K of "multimorbidity features", each associated with a morbidity profile ρ_k, $k = 1, \ldots, K$. Each profile describes the probabilities of each morbidity occurring (Figure 10A). The *prevalence* of each feature, that is the probability that an individual possesses the feature, is time-varying. This allows some features to be associated with increasing age, while others may only be linked to younger individuals or to transient periods in an adult life (Figure 10B). Each individual in the cohort can possess any number of these multimorbidity features. This feature allocation approach means an individual is not constrained to belong only to a single feature (Figure 10C) and that different types of multimorbidity can appear and disappear throughout their lifetime. Finally, the observed data consists of individuals, labelled by age, and an indicator for the presence of each morbidity (Figure 10D). Our objective is to perform inference to learn feature morbidity profiles and prevalences from the observed data.

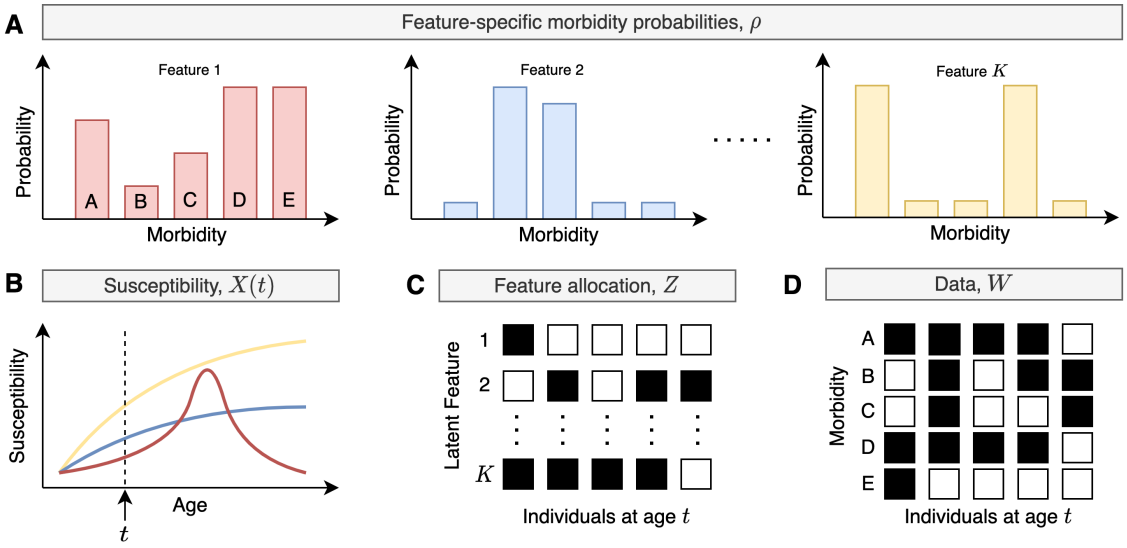

Figure 10: **Proposed Model.** (A) The model assumes the existence of up to K latent multimorbidity features each of which is associated with a profile of morbidity probabilities. Here, feature 1 has high probabilities of morbidities A, D, E, while feature 2 has high probabilities associated with morbidity B and C. (B) Each feature is associated with a time-varying prevalence. For example, feature 1 is prevalent around a certain range in middle-age, while feature 2 increases with age. (C) Each individual may possess any of the K features, hence this is a latent feature allocation model. (D) Finally, the observed data consists only of the morbidities recorded for each individual and their ages at the time of the study when the data was collected.