Latent Variable Approaches to Multimorbidity Analysis

Woojung Kim
University of Warwick

(with Paul Jenkins, University of Warwick and Christopher Yau, University of Oxford)

Background

Motivating Example: Multimorbidity

The presence of **two or more** (chronic) disease conditions in the same individual.

Increasing prevalence due to aging populations.

Complex diagnostic and treatment regimes.

Most clinicians are trained to treat or manage single conditions.

Major burden on health services.

Introducing “Mary”

Mary is a 72 year old female, she has:

- Diabetes,
- Rheumatoid arthritis,
- Chronic Obstructive Pulmonary Disease (COPD),
- Depression.

She is currently taking 11 medications and describes her health situation as “an endless struggle” with prolonged clinical consultations.

What is the next best treatment for her diabetes, *in the context of her RA, COPD and depression*?

Existing works

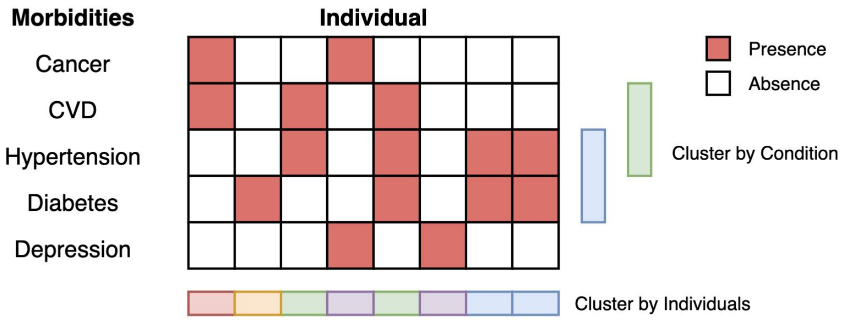
Q. Which conditions co-occur together?

Studies have used electronic health records (EHRs) or cohort data collections to understand pattern of disease co-occurrence.

Many studies have identified multimorbidity clusters, groups of diseases that commonly occur together.

Common analysis techniques include

- K-Means
- Hierarchical Clustering Analysis
- Latent class analysis,



Morbidities	**Individual**							
Cancer	🟥	□	□	🟥	□	□	□	□
CVD	🟥	□	🟥	□	🟥	□	□	□
Hypertension	□	□	🟥	□	🟥	□	🟥	🟥
Diabetes	□	🟥	□	□	🟥	□	🟥	🟥
Depression	□	□	□	🟥	□	🟥	□	□

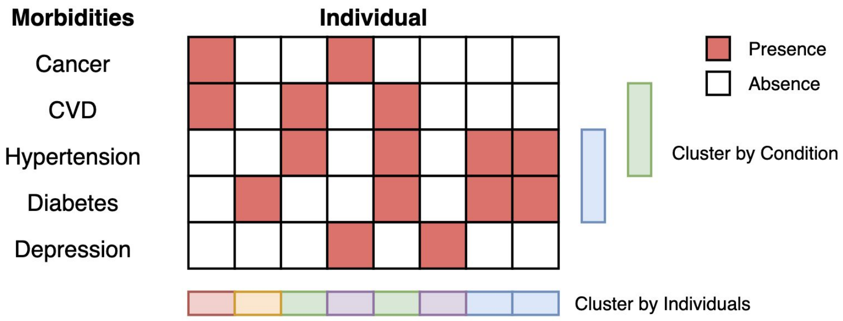
Multimorbidity Analysis

Q. Which conditions co-occur together?

- Cluster by conditions

Q. What are the conditions with the most significant health impacts?

- Cluster by individuals



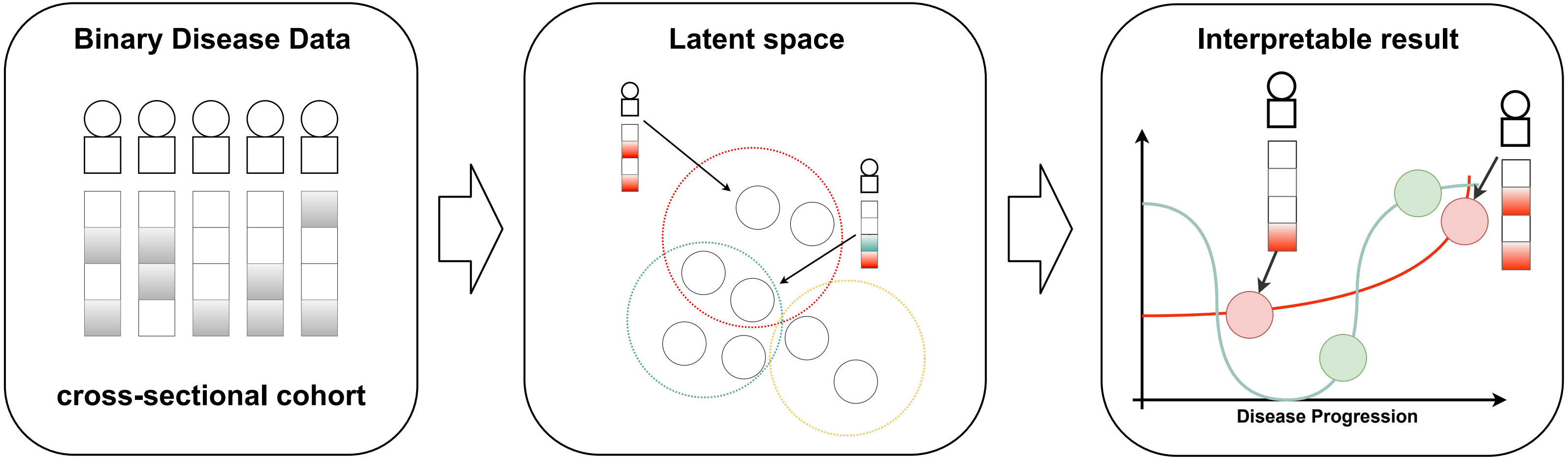
Morbidities	Individual							
Cancer	🟥			🟥				
CVD	🟥		🟥		🟥			
Hypertension			🟥		🟥		🟥	🟥
Diabetes		🟥			🟥		🟥	🟥
Depression				🟥		🟥		

🟥 Presence

□ Absence

Unsupervised Learning

Unsupervised learning of low-dimensional structure from high-dimensional health-related attributes



Binary Disease Data

cross-sectional cohort

Latent space

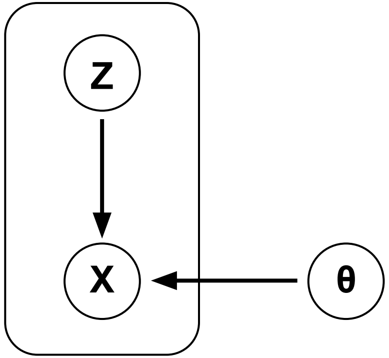
Interpretable result

Goal is to obtain latent representations that allows to identify patterns of multimorbidities

Latent Variable Approaches to Multimorbidity Analysis

Notation

- **x** - high-dimensional health-related attributes
- **θ** - a set of parameters associated with models
- **z** - low-dimensional (latent) variables



Bayesian Hierarchical Model:

$$\mathbf{z}_i \sim p(\mathbf{z}), i = 1, \cdots, N$$

$$\mathbf{x}_i \sim p_\theta(\mathbf{x}|\mathbf{z}), i = 1, \cdots, N$$

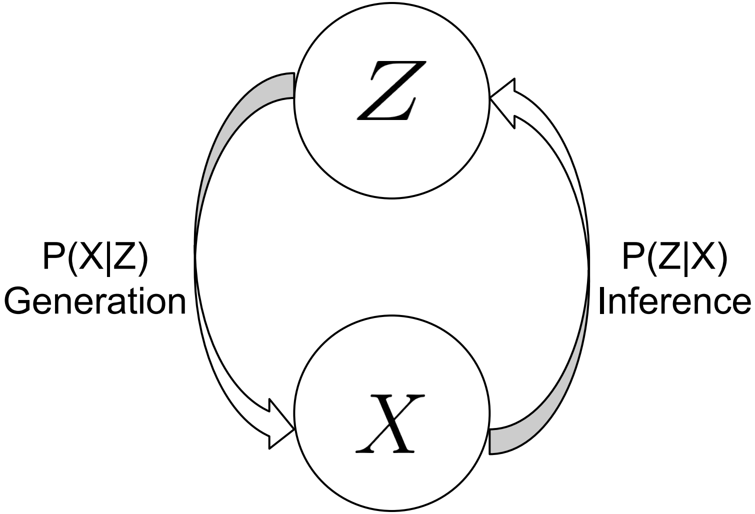
We are interested in discovering some (interesting) low-dimensional representation **z** from data **x**

Bayesian Inference

- Inference about the unknown is through the posterior distribution

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

- For most interesting models, the denominator is often intractable
- We appeal to approximate posterior inference including
  - MCMC: slow but accurate
  - Variational inference: scalable but less accurate

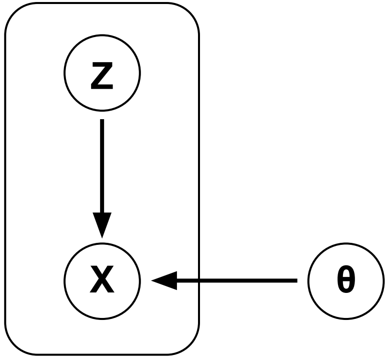


Wright-Fisher Multimorbidity Trajectory Model

x - Individual’s binary morbidity indicators

θ - parameters associated with the model

z - Discrete latent variables indicating multimorbidity membership of each morbidity



Goal is to identify clusters of morbidities and their prevalence trajectory over time.

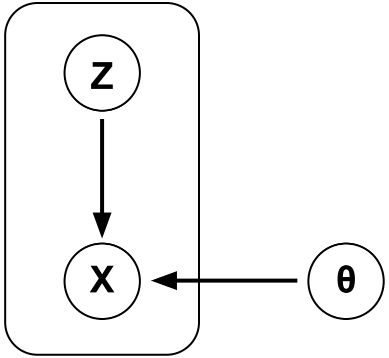
Multimodality Multimorbidity Variational Autoencoder

x - Mixed-type health-related attributes:

- binary morbidity indicators
- continuous physiological measurements
- survival outcome

θ - parameters associated with the model

z - continuous latent variables that act as latent health summary of individual



Goal is to identify (low-dim) continuous representations (Z) that summarize individual’s health

Feature Allocation Approach for Multimorbidity Trajectory Modelling

Woojung Kim
University of Warwick and Alan Turing Institute

(with Paul Jenkins, University of Warwick and Christopher Yau, University of Oxford)

Existing works

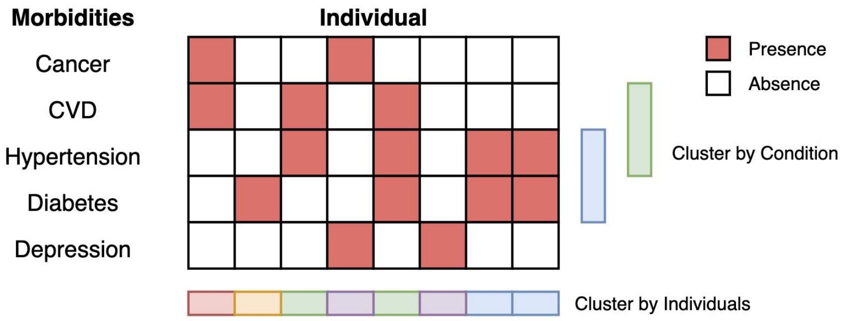
Q. Which conditions co-occur together?

Studies have used electronic health records (EHRs) or cohort data collections (e.g., *cross-sectional study*) to understand pattern of disease co-occurrence.

Many studies have identified multimorbidity clusters, groups of diseases that commonly occur together.

Common analysis techniques include

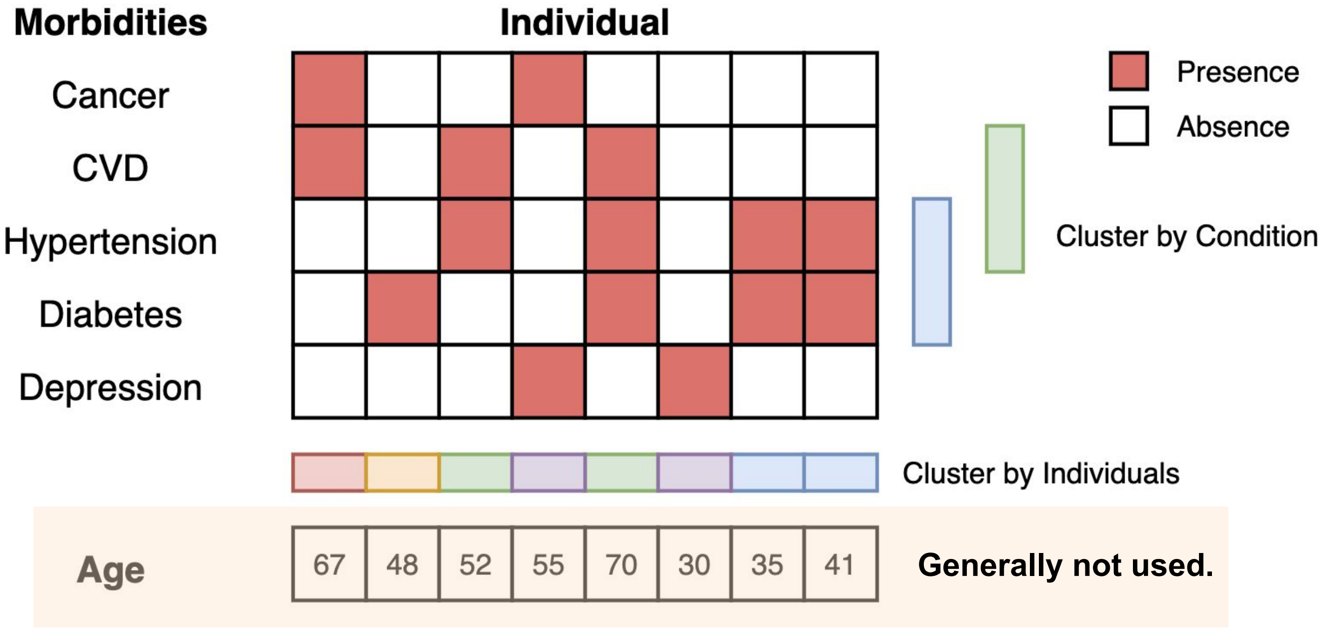
- K-Means
- Hierarchical Clustering Analysis
- Latent class analysis,



Morbidities	Individual							
Cancer	🟥	□	□	🟥	□	□	□	□
CVD	🟥	□	🟥	□	🟥	□	□	□
Hypertension	□	□	🟥	□	🟥	□	🟥	🟥
Diabetes	□	🟥	□	□	🟥	□	🟥	🟥
Depression	□	□	□	🟥	□	🟥	□	□

🟥 Presence
□ Absence

Limitation of existing work



Morbidities	Individual							
Cancer	🟥	□	□	🟥	□	□	□	□
CVD	🟥	□	🟥	□	🟥	□	□	□
Hypertension	□	□	🟥	□	🟥	□	🟥	🟥
Diabetes	□	🟥	□	□	🟥	□	🟥	🟥
Depression	□	□	□	🟥	□	🟥	□	□
Age	67	48	52	55	70	30	35	41

🟥 Presence

□ Absence

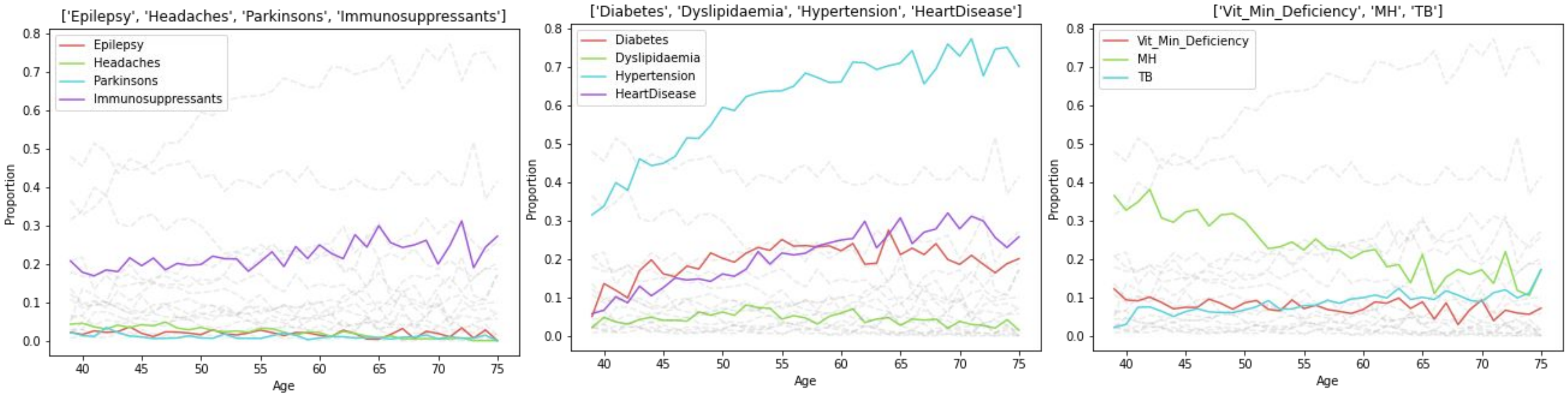
Cluster by Condition

Cluster by Individuals

Generally not used.

Empirical age dependence trends in disease prevalence

Clear age dependencies in the empirical prevalence of common, single conditions:

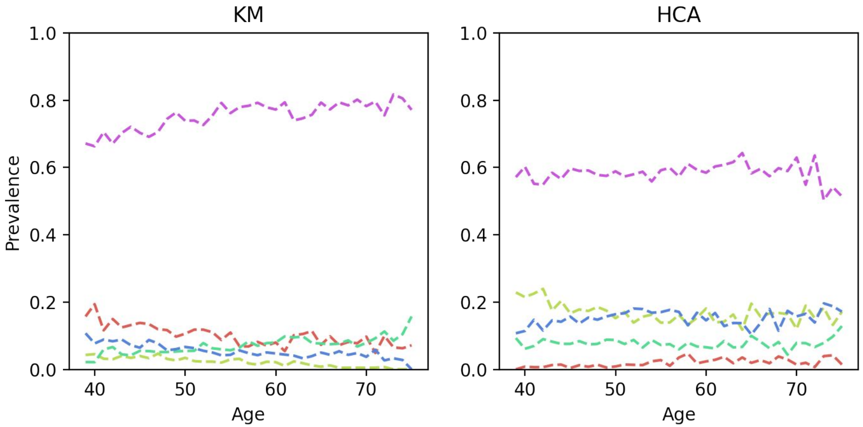


Age is often either (i) not used(/available) or (ii) used to stratify the population into groups.

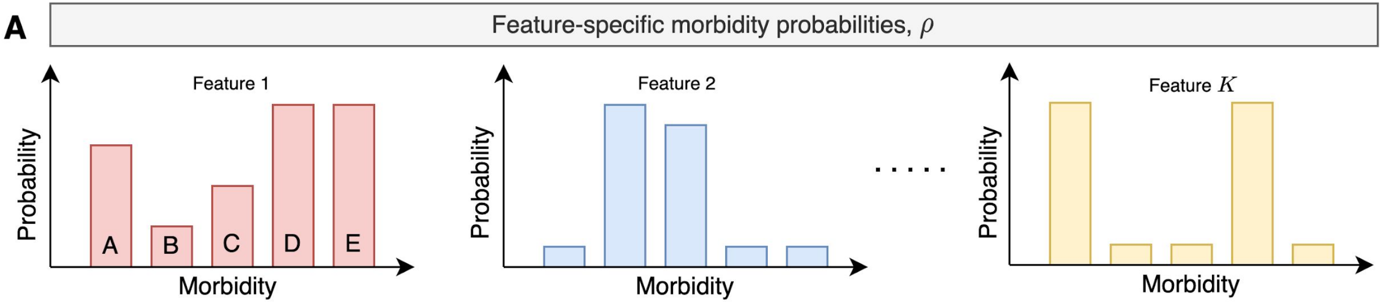
The utility of clusters

Clusters are often dominated by a single condition (or very closely related conditions) or involve clusters of very common disease.

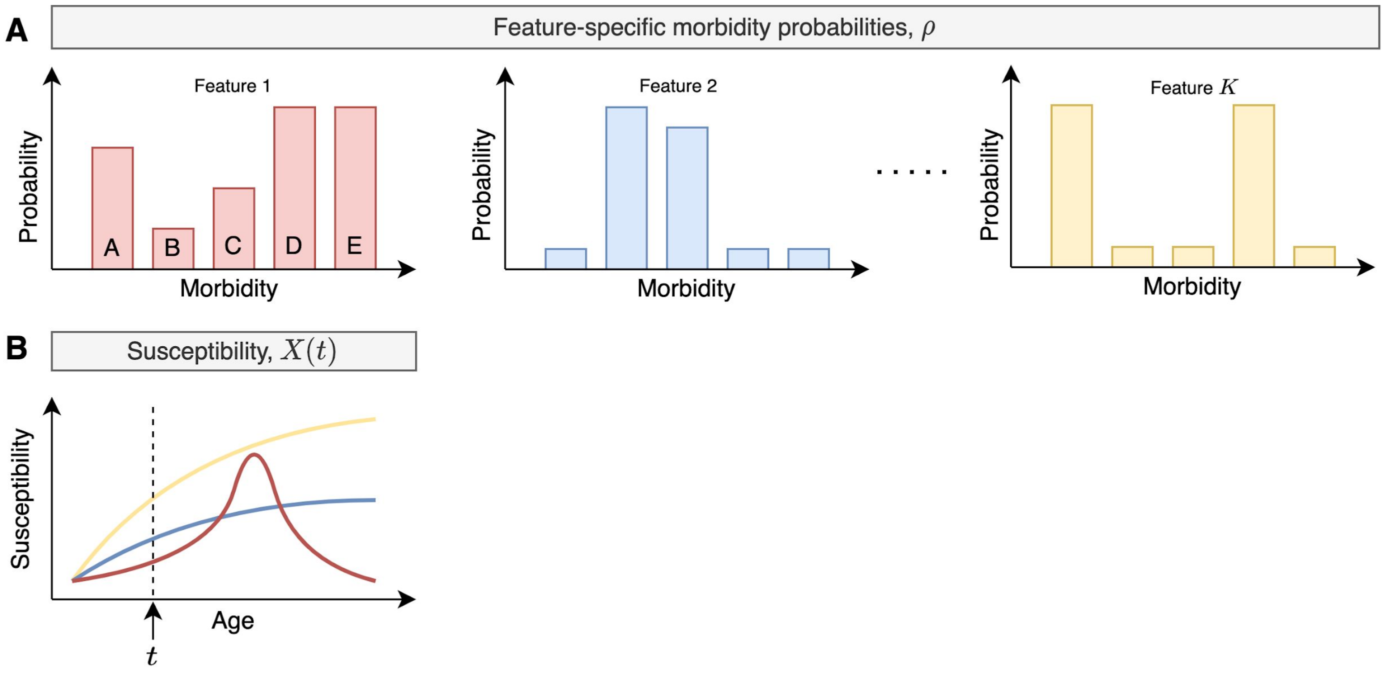
Clinicians often question the value of these disease clusters (“*we know these already*”, “*what do we do with them*?”)



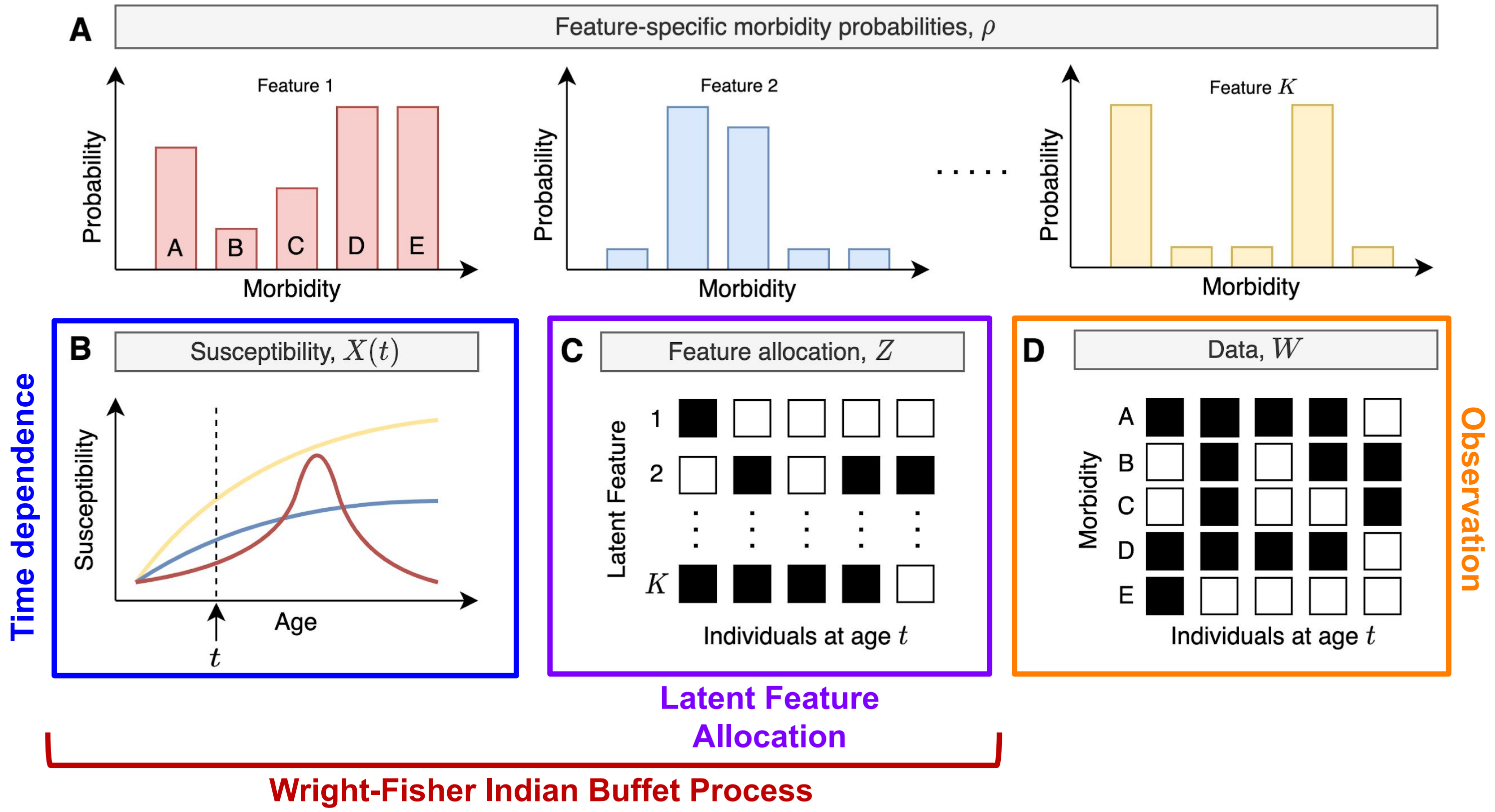
Time-dependent Latent Feature Allocation Model



Time-dependent Latent Feature Allocation Model



Time-dependent Latent Feature Allocation Model



1	■	□	□	□	□
2	□	■	□	■	■
	⋮	⋮	⋮	⋮	⋮
K	■	■	■	■	□

A	■	■	■	■	□
B	□	■	□	■	■
C	□	■	□	□	■
D	■	■	■	■	□
E	■	□	□	□	□

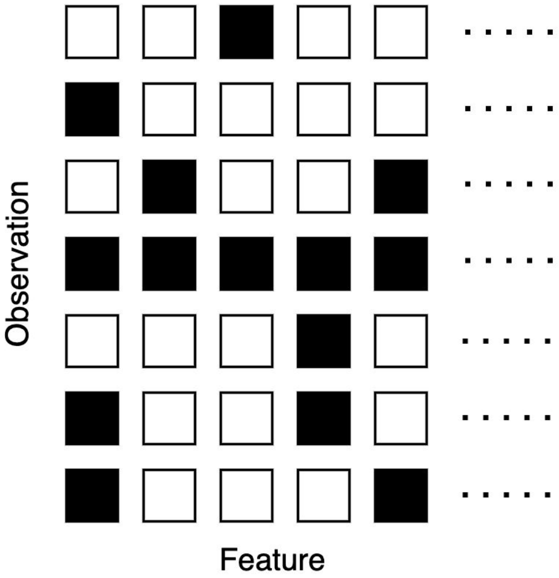
Background: Indian Buffet Process (IBP)

- Indian Buffet Process (IBP) is a distribution over binary matrices with unbounded features.
- Widely used as a prior for models with potentially infinite number of latent features.
- To generate a random binary matrix from IBP, consider the beta-binomial model:

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

$$Z_{ik} | X_k \sim \mathrm{Bernoulli}(X_k)$$

This is pre-limiting model of the two-parameter IBP ($K \to \infty$).



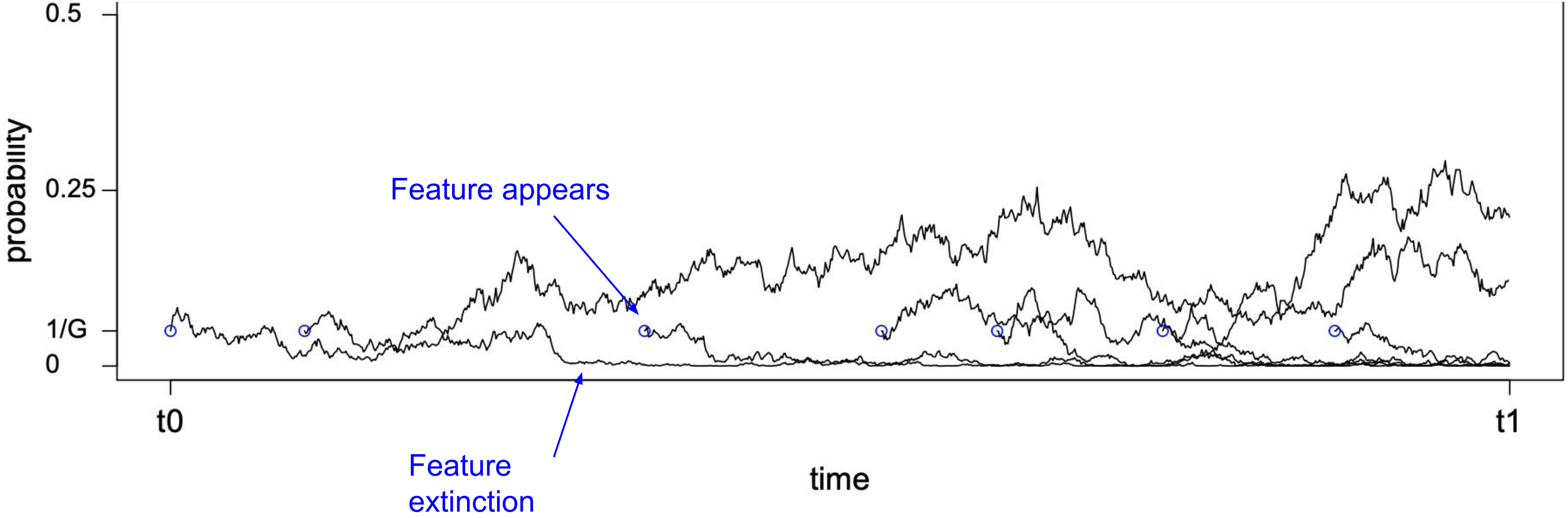
Note - number of features are unknown

Background: Dependent IBP

We would like to make the model time-dependent by letting the probability $X_k(t)$ of each feature evolve.

How to evolve?

We want to allow features to become extinct (multimorbidity clusters disappear) or appear at random times (multimorbidities emerge).



Background: Wright-Fisher diffusion

For each feature, consider $\mathrm{Beta}\left(\frac{\alpha\beta}{K}, \beta\right)$ as the stationary distribution of a W-F diffusion with $\frac{\alpha\beta}{K} > 0$ and $\beta > 0$.

Wright-Fisher diffusion model:

- the continuous time limit of the W-F model,
- genetic evolution in fixed-size population,
- permits appearance of “mutations” (new features),
- permits extinction of “mutations” (disused features),

Transition function not available in closed form but exact simulation from the W-F diffusion process is possible (Jenkins & Spano, 2017).

Background: Wright-Fisher IBP

To generate a random binary matrix $N \times K$ from WF-IBP:

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

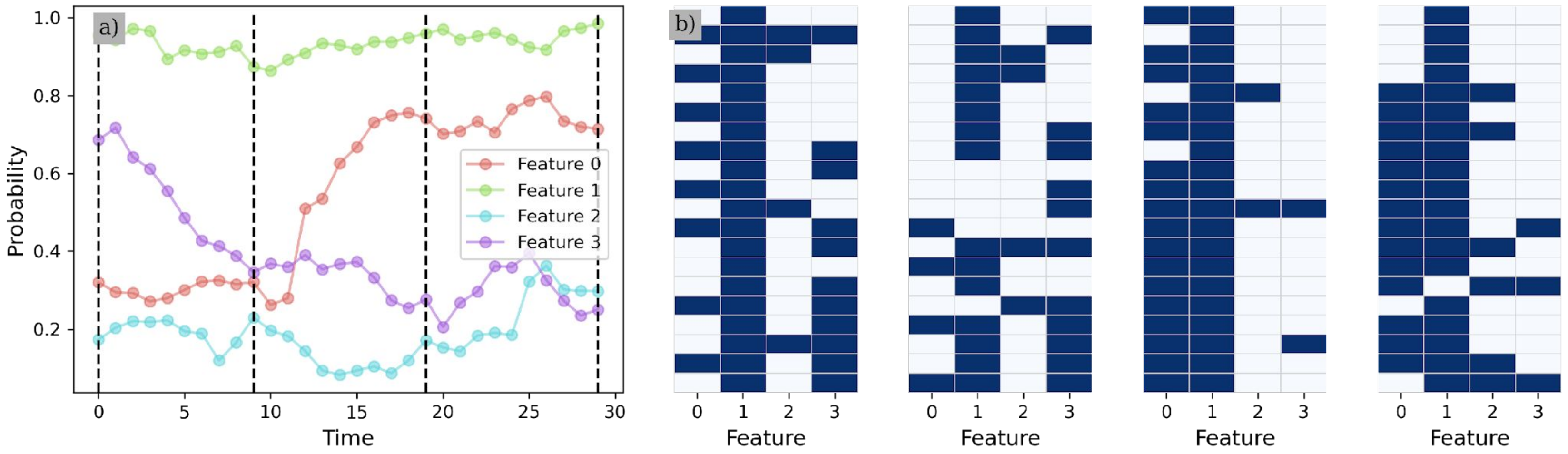
$$z_{ikt} | X_k(t) \sim \mathrm{Bernoulli}(X_k(t)), \quad \forall i$$

At any given t, matrix Z_{ik} corresponds to a sample from an IBP.

In the finite-K approximation, as the W-F diffusion starts at stationarity, this construction coincides marginally with the beta-binomial model.

The parameters of the W-F diffusion are positive, so that neither fixation nor absorption ever occurs and the number K of features remains constant.

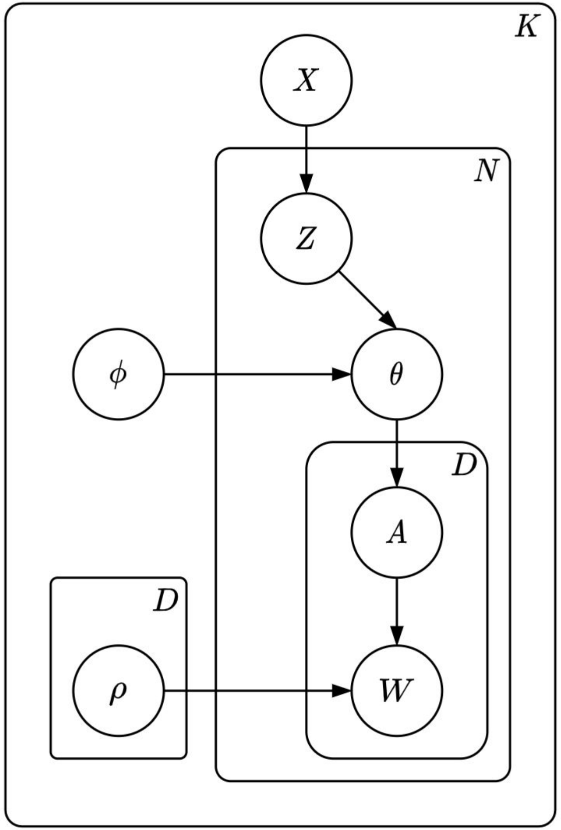
Background: Wright-Fisher IBP



(a) Feature probabilities evolve across time according to WF diffusion; (b) Feature allocation matrices at times shown in black dashed line in (a)

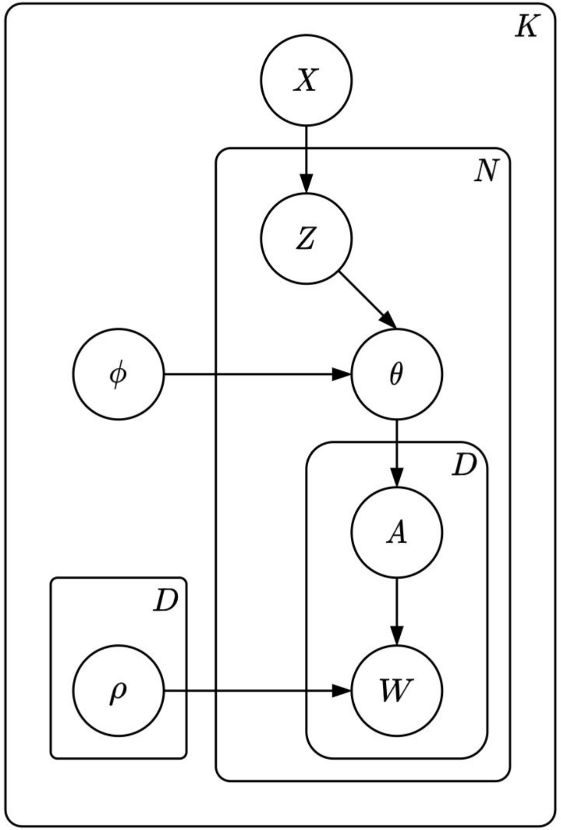
Notations

- K - # of latent features (i.e., multimorbidity)
- N - # of individuals
- D - # of morbidities
- T - # of age groups
- (i, t) : individual indice (cross-sectional dataset)



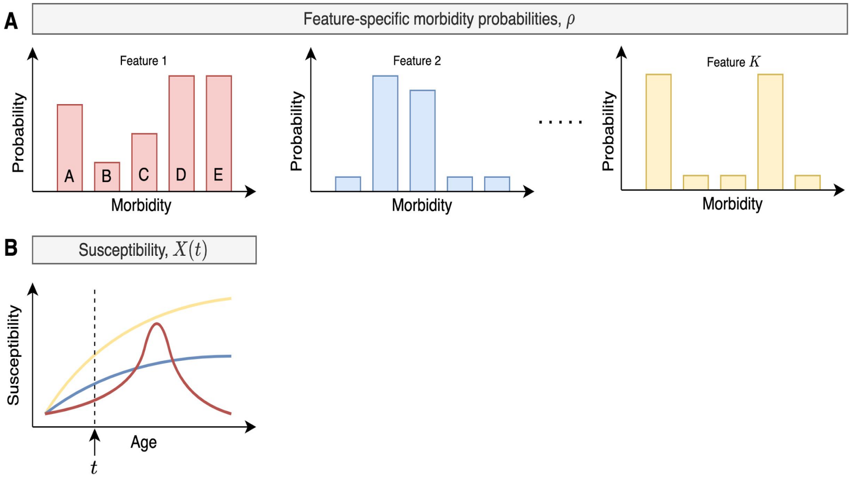
Notations

- $W_{itd} \in \{0, 1\}$ - binary morbidity presence observation data
- $X_k(t)$ - probability of having kth multimorbidity feature at time t
- $Z_{itk} \in \{0, 1\}$ - indicator that the i-th individual belongs to multimorbidity feature k at time t
- ρ_{kd} - feature-specific morbidity probabilities



Goal is to identify clusters of morbidities (ρ) and their prevalence trajectory over time (X).

Generative Process



Step 1: Sampling multimorbidity feature properties

For each feature $k \in \{1, \cdots, K\}$:

1. Sample feature-specific susceptibility:

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

2. Sample feature-specific disease probabilities:

$$\rho_{kd} \sim \mathrm{Beta}(\eta_0, \eta_1).$$

3. Sample feature intensity:

$$\phi_k \sim \mathrm{Gamma}(\gamma, 1).$$

Generative Process

Step 2: Sampling patient-specific properties

For each time $t \in \{1, \cdots, T\}$ and each patient i:

- Sample the kth feature presence indicator:

$$z_{itk} \sim \mathrm{Bernoulli}(x_{kt})$$

- Sample the patient-specific scaled-feature probabilities:

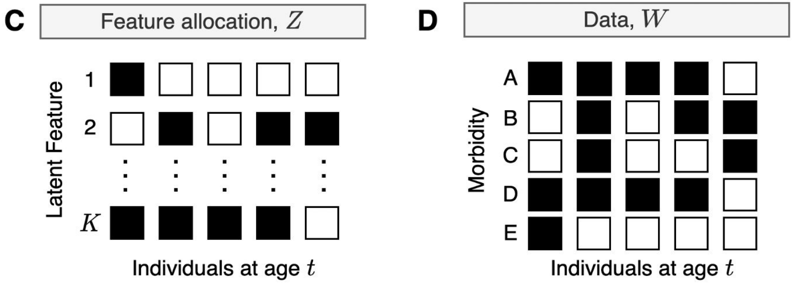
$$\theta_{it} \sim \mathrm{Dirichlet}(z_{it} \cdot \phi)$$

- For each disease, $d \in \{1, \cdots, D\}$:
  1. Sample the cluster from which the disease originates:

$$a_{idt} \sim \mathrm{Categorical}(\theta_{it})$$

  2. Sample the observed disease status:

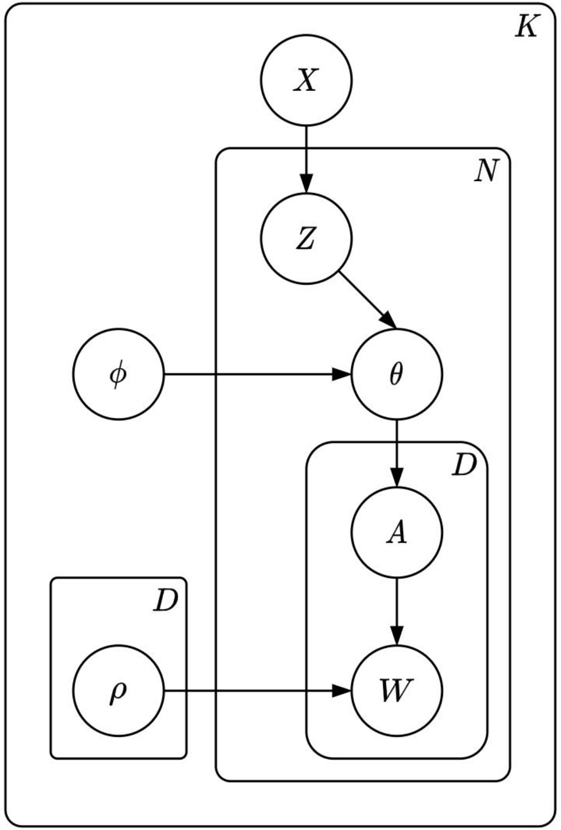
$$w_{idt} \sim \mathrm{Bernoulli}(\rho_{a_{idt} d})$$



Posterior Inference

We can largely integrate out θ, ρ and draw posterior sample of X, Z, ϕ, A:

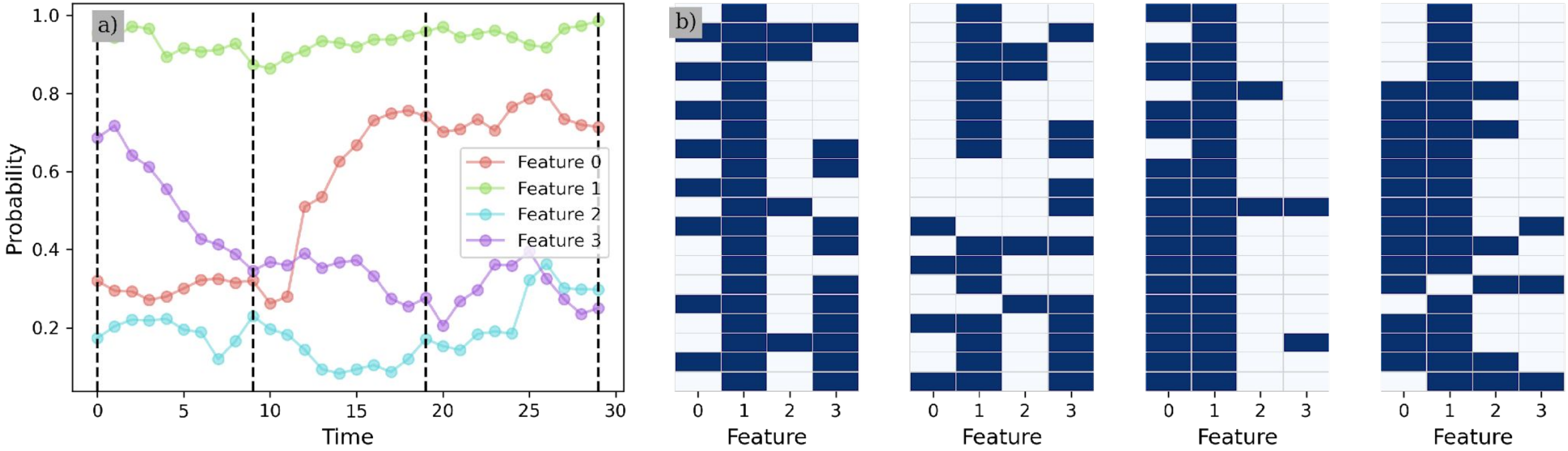
1. $X \sim X|Z$: Sequential Monte Carlo
2. $Z \sim Z|A, X, \phi$: Hamming Ball Sampler
3. $\phi \sim \phi|Z, A$: Metropolis-Hastings
4. $A \sim A|W, Z, \phi$: Gibbs



Posterior Inference of feature probabilities x

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

$$z_{ikt} | X_k(t) \sim \mathrm{Bernoulli}\left(X_k(t)\right), \quad \forall i$$



SMC: Posterior sampling of feature probabilities X

The prior for each feature is $\mathrm{Beta}\left(\frac{\alpha\beta}{K}, \beta\right)$ and the column-wise sums of Z_{t_0} are realisations from binomials, by conjugacy:

$$X_k(t_0)|Z_{t_0} \sim \mathrm{Beta}\left(\frac{\alpha\beta}{K} + n_{k,t_0}, \beta + N_{t_0} - n_{k,t_0}\right),\ k = 1, \dots, K,$$

and

$$n_{k,t} = \sum_{i=1}^{N_t} z_{ikt}$$

is the number of individuals in matrix Z_t possessing feature k.

SMC: Posterior sampling of feature probabilities

1. Draw a number of particles from:

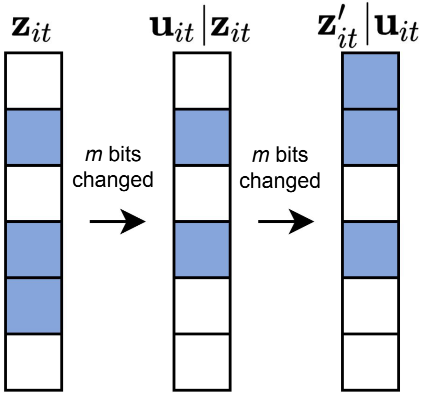
$$X_k(t_0)|Z_{t_0} \sim \mathrm{Beta}\left(\frac{\alpha\beta}{K} + n_{k,t_0}, \beta + N_{t_0} - n_{k,t_0}\right), \; k = 1, \ldots, K,$$

2. Propagate each particle forward in time according to $X_k(t_1)|X_k(t_0) \sim \mathrm{WF}\left(\frac{\alpha\beta}{K}, \beta\right)$.
3. Weight according to number of binomial likelihood of seeing n_t individuals out of N_t in Z_t with the feature active, $x_k(t)^{n_{kt}}(1 - x_k(t))^{N_t - n_{kt}}$.
4. Sample weighted particles with replacement and propagate.

Posterior sampling of latent feature allocations Z

Conditional posterior:

$$P(\mathbf{Z_{it}}|\mathbf{n_{it}}, \mathbf{X}(t), \phi) \propto \binom{D}{n_{its_1} \cdots n_{its_S}} \frac{B(\phi_{s_1} + n_{its_1}, \cdots, \phi_{s_S} + n_{its_S})}{B(\phi_{s_1}, \cdots, \phi_{s_S})} \prod_{k=1}^{K} X_k(t)^{Z_{itk}} (1 - X_k(t))^{1-Z_{itk}}$$


$\mathbf{z}_{it}$
$\mathbf{u}_{it}|\mathbf{z}_{it}$
$\mathbf{z}'_{it}|\mathbf{u}_{it}$
m bits changed
m bits changed

Ideally, we want to jointly update all feature allocations for each individual at each time.

Adopt Hamming Ball sampling approach (Titsias & Yau, 2017) to avoid exhaustive enumeration over all 2^K possibilities.

Baseline Method

- **IBP (Ruiz et al, 2014)**
    - Indian Buffet Process without temporal dependence
        - allows for multiple comorbidity membership

- **Latent Factor Analysis (Linzer and Lewis 2011)**
    - Latent variable model for clustering
        - An individual can be influenced by only one comorbidity

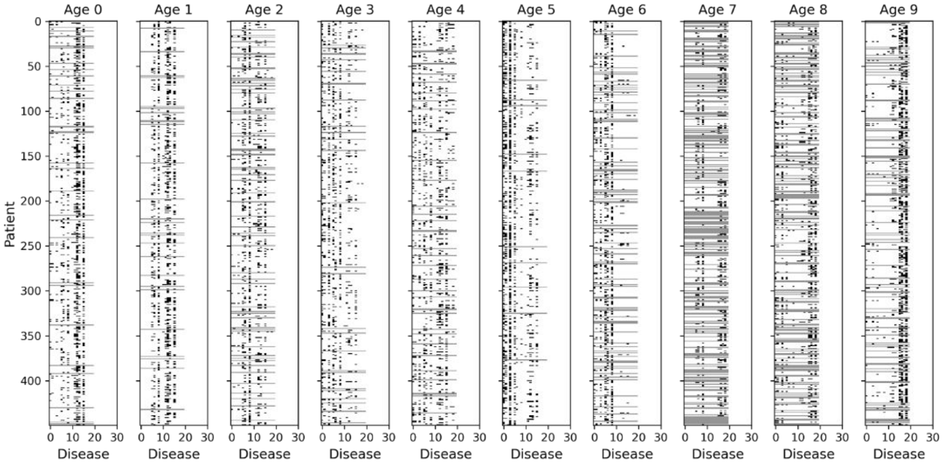
Simulation Experiment I

N = 900, T = 6, K = 3, D = 20

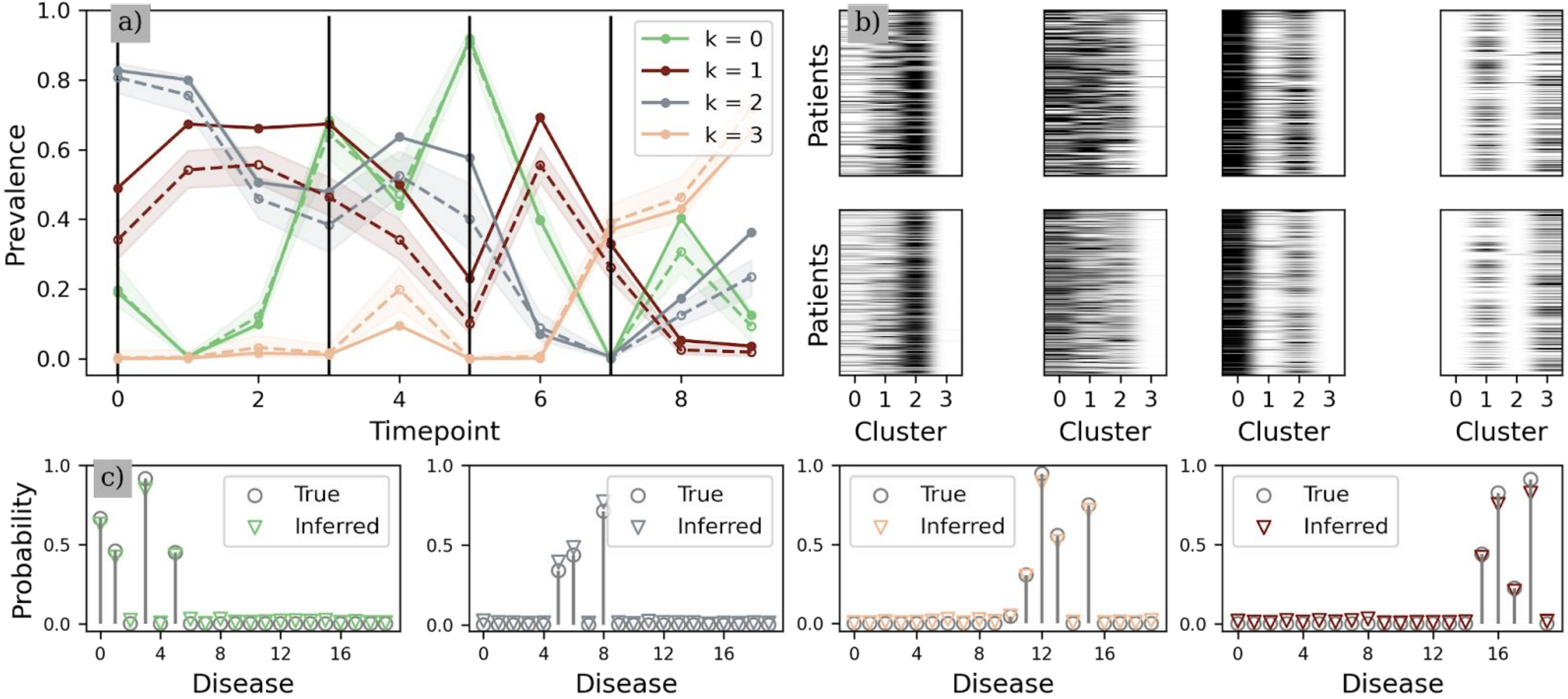
Each pair of clusters share a common disease – simulating the situations where there are highly prevalent disease (e.g. depression).

Add noise to these observations by flipping each of their disease indicators from 0 to 1 and 1 to 0 with probability 0.01 and 0.05 respectively.

MCMC for 5000 iterations with 4000 burn-ins.

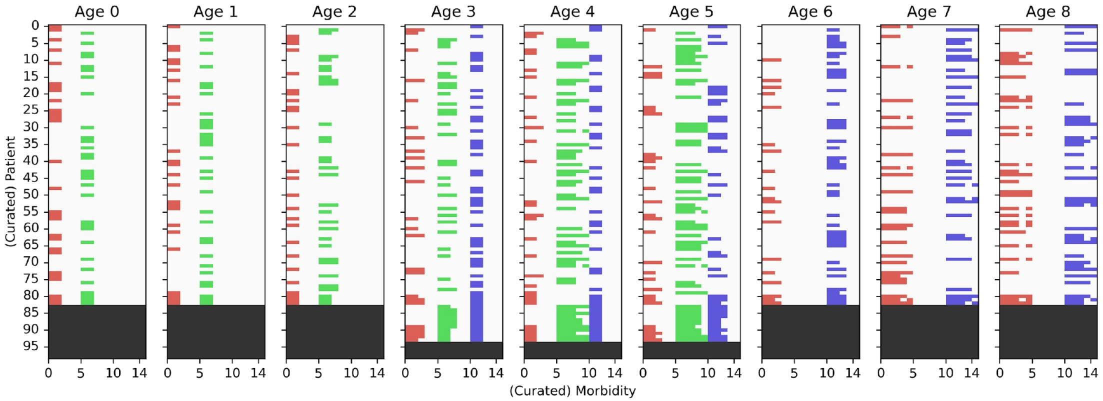


Simulation Experiment I



Simulation Experiment II

Curated toy dataset – not from the generative model.

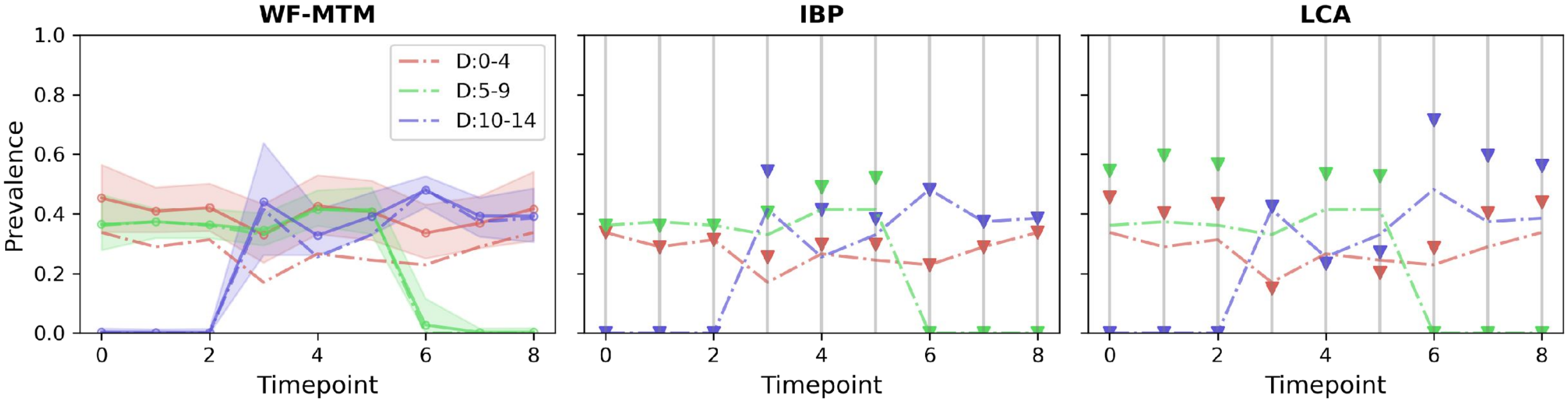


Conditions 0-4 remains throughout
Conditions 5-19 appear in early age only
Conditions 11-14 appear in middle age

Variable number of individuals at each age simulating a real cross-sectional cohort.

All cluster's disease profile changes over time.

Simulation Experiment II



(a) Empirical proportion of each cluster is closely followed by the feature trajectory from WF-MTM; (b) other clustering approaches, e.g. LCA, KM and HCA, continue to over-estimate the ground truth

Case study: Iran

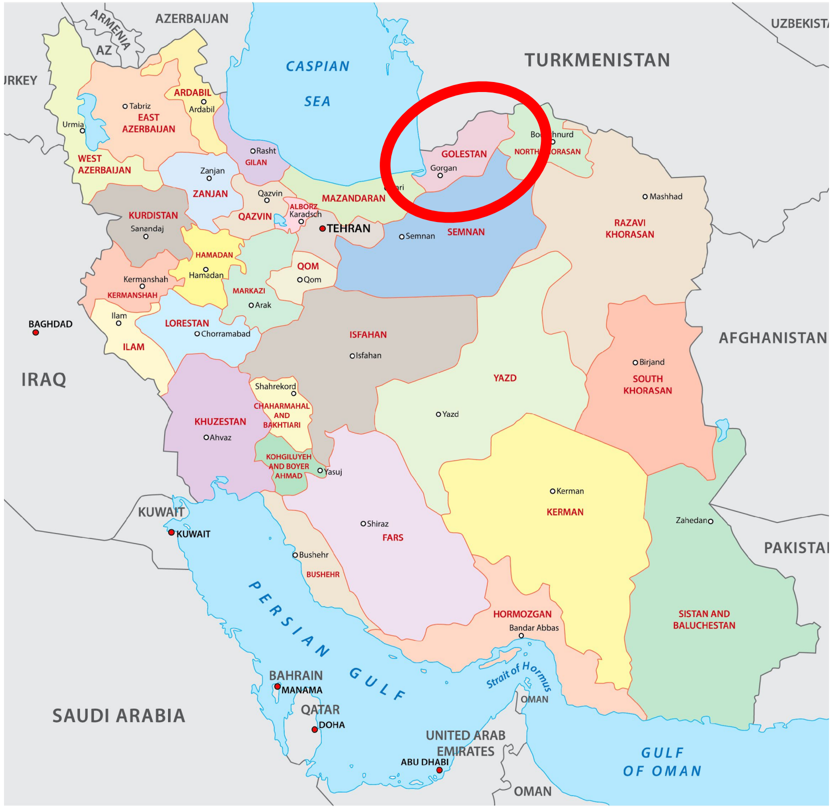
54,000 participants from the Golestan cohort (North East of Iran).

Up to 14 years follow up.

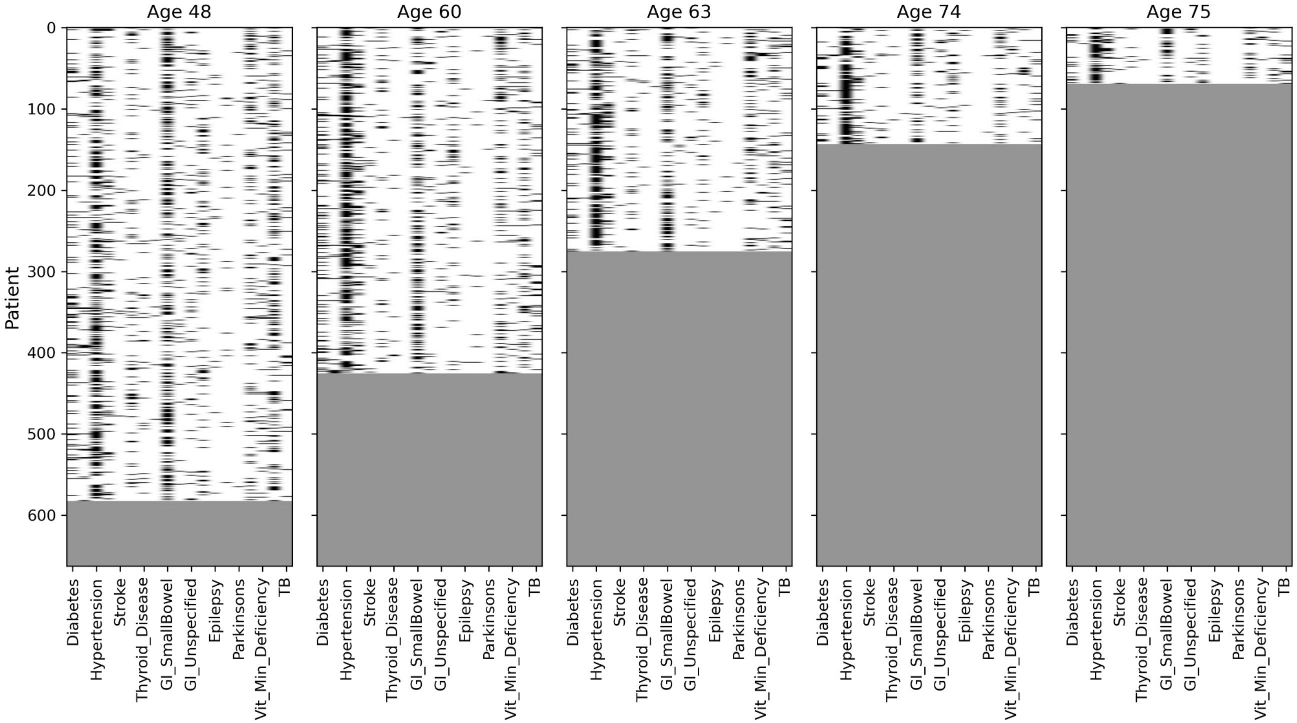
Disease status for up to 30 conditions (removed very low or very high-frequency conditions).

Individual-level data including socioeconomic, clinical and prescription data.

Relatively “complete” at baseline but follow-up is patchy.

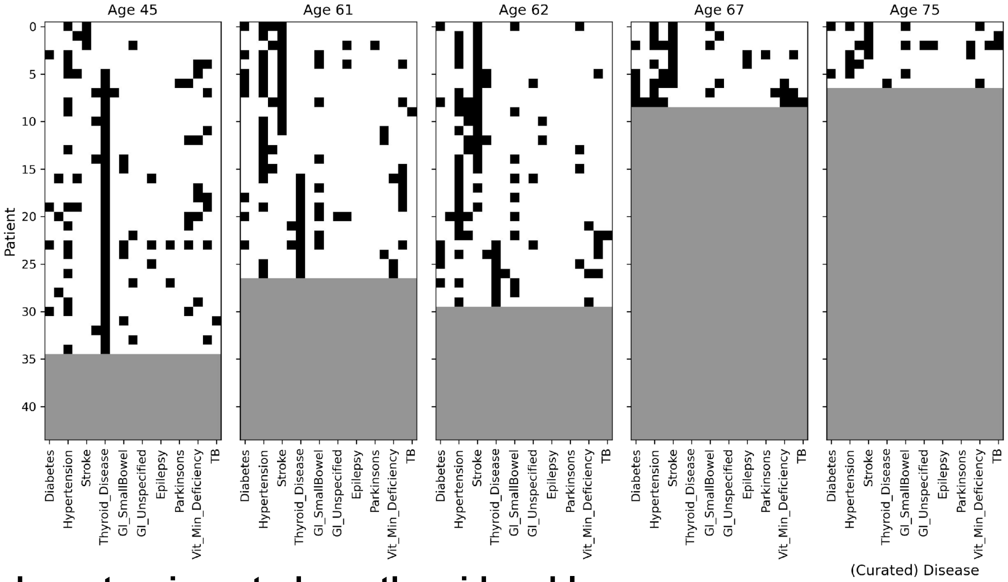


Golestan Cohort Study: (Sparse) Binary Data



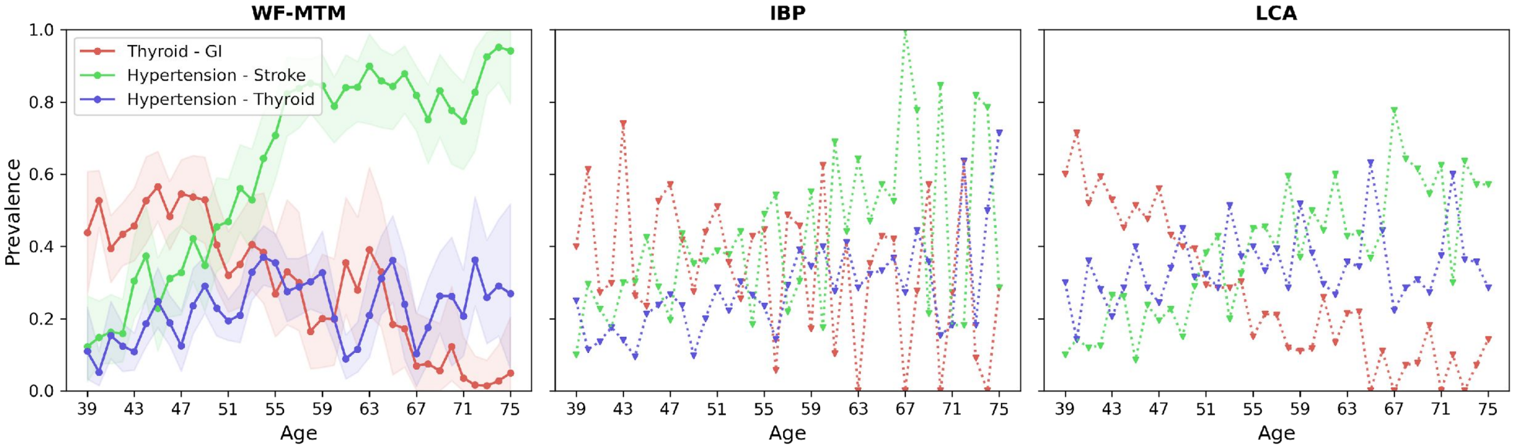
Experiment III

Semi-curated dataset – subset of real-world dataset



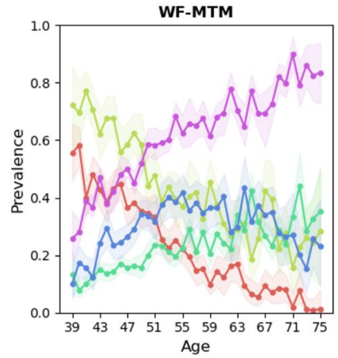
Each patient has either hypertension, stroke or thyroid problems.

Experiment III



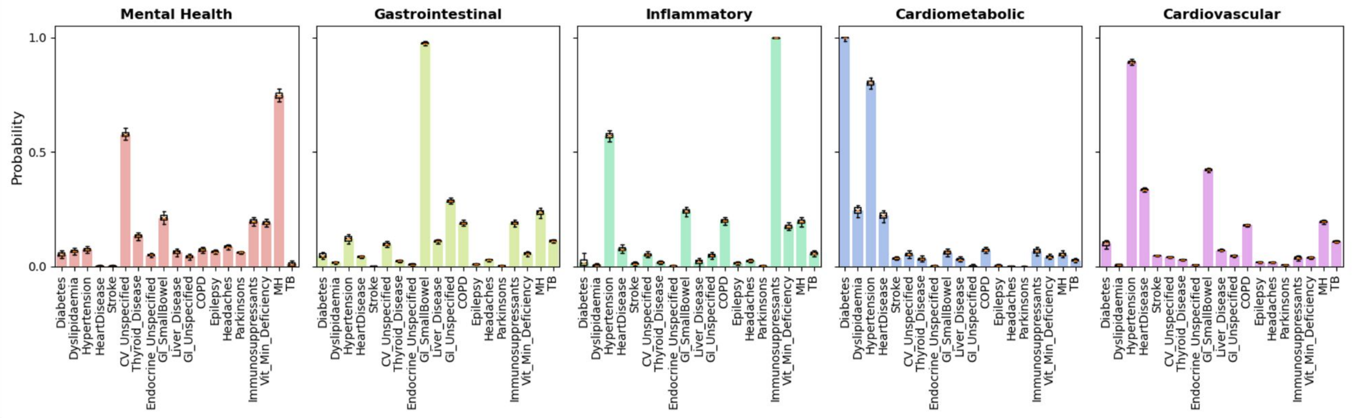
(a) These methods were able to summarise dominant multimorbidity patterns in the form of separate latent features (clusters); (b) WF-MTM stood out by its ability to recapitulate a clear age-linked dependencies.

Golestan Cohort Study

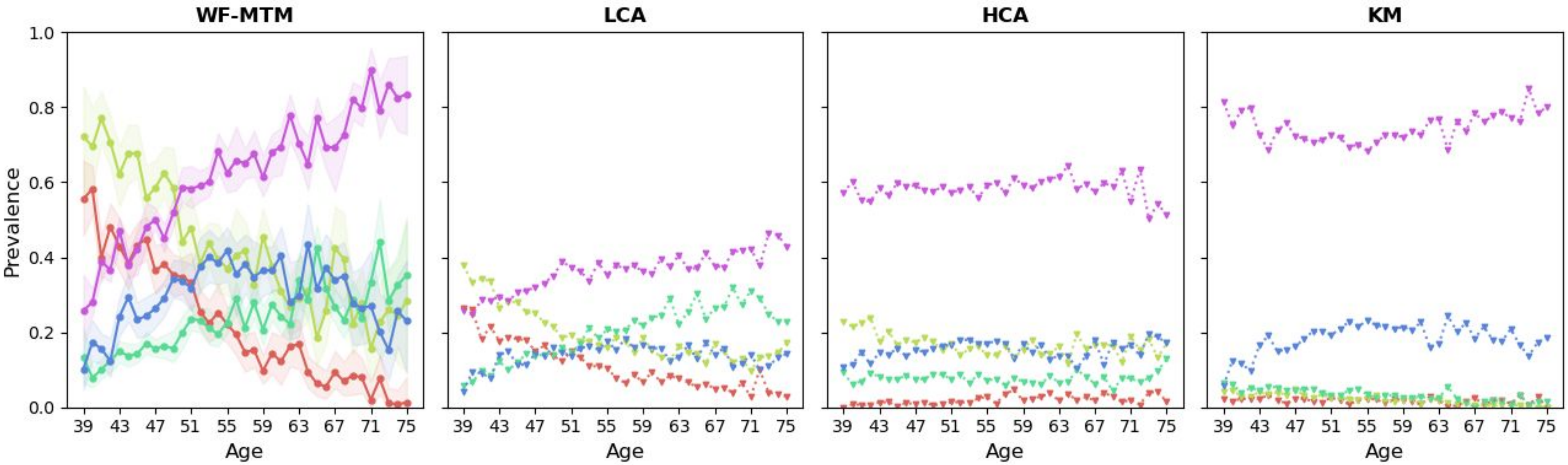


Incorporating age allows WF-MTM to identify time-dependent emergences of morbidity profiles

- ○ Increased mental health conditions in middle ages
- ○ Increasing imflammatory and cardiovascular disease with age

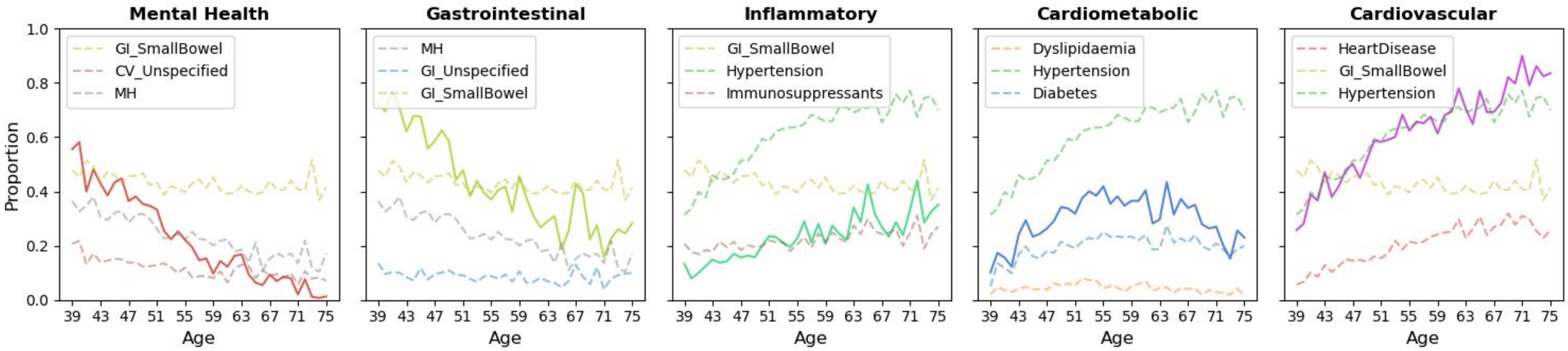


Performance evaluation I



WF-MTM features exhibit **clear** age-dependency in terms of their prevalence.

Performance evaluation II

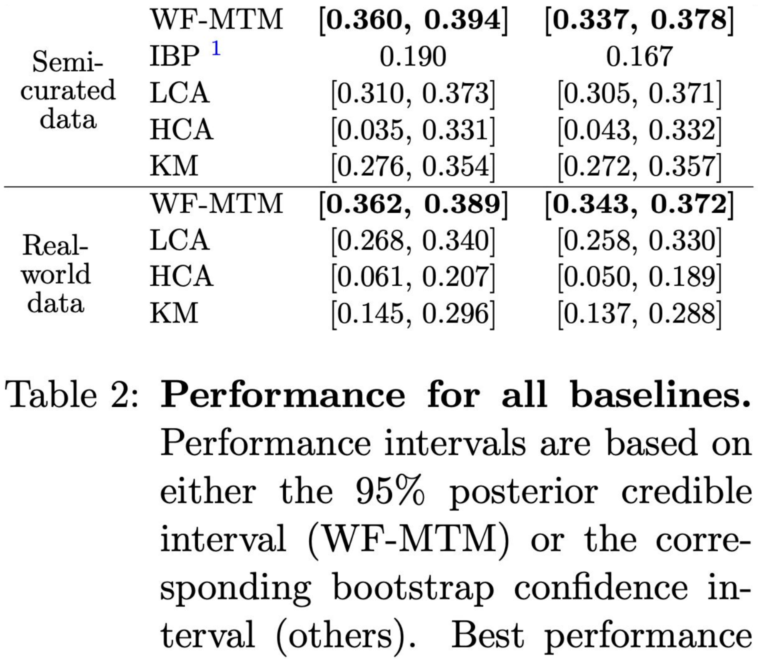


The reconstructed temporal prevalences from WF-MTM are consistent with that of its leading conditions

Performance evaluation III

- Quantitative assessment

  Weighted average of correlations between the (estimated) temporal trajectory of a feature and that of each single condition, weighted by the corresponding morbidity profile



Semi-curated data	WF-MTM	**[0.360, 0.394]**	**[0.337, 0.378]**
	IBP [1]	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

WF-MTM features exhibit **clear** age-dependency in terms of their prevalence.

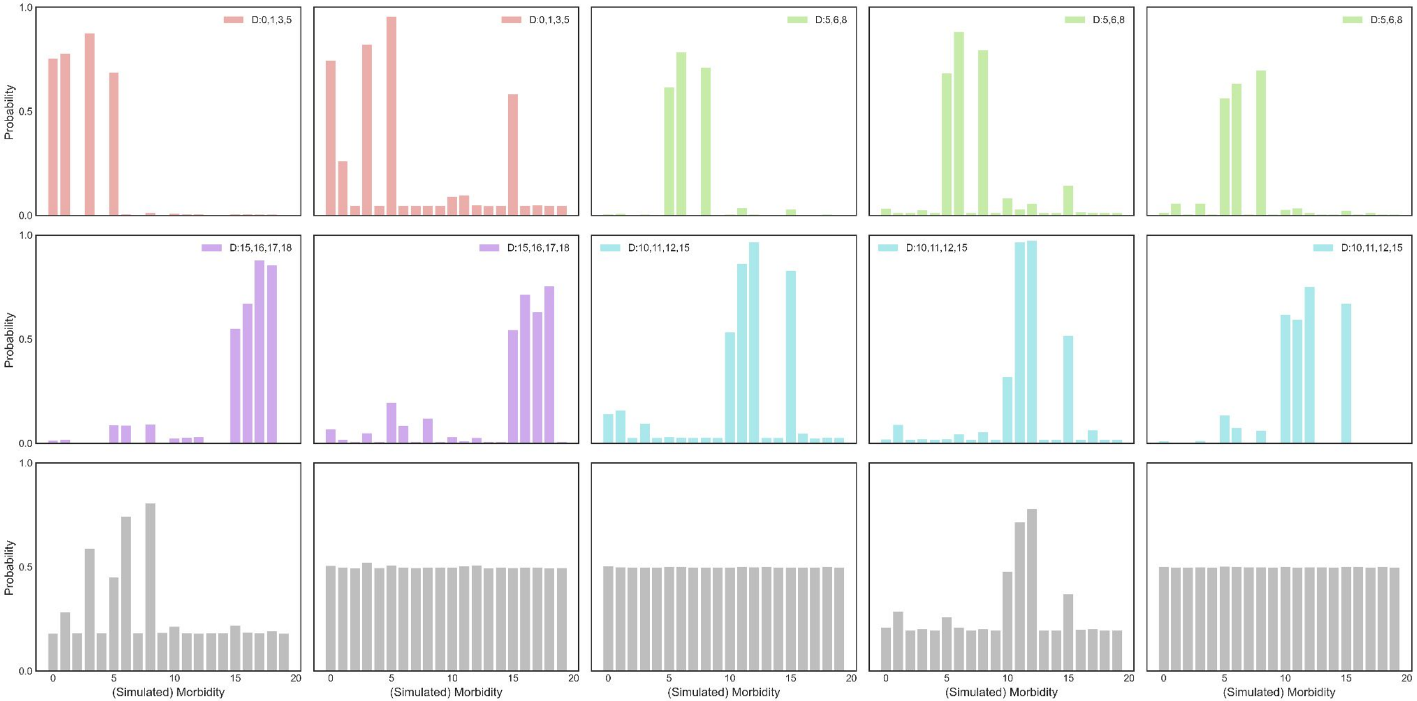
Appendix: Model misspecifications

- The number of latent features K is a hyperparameter

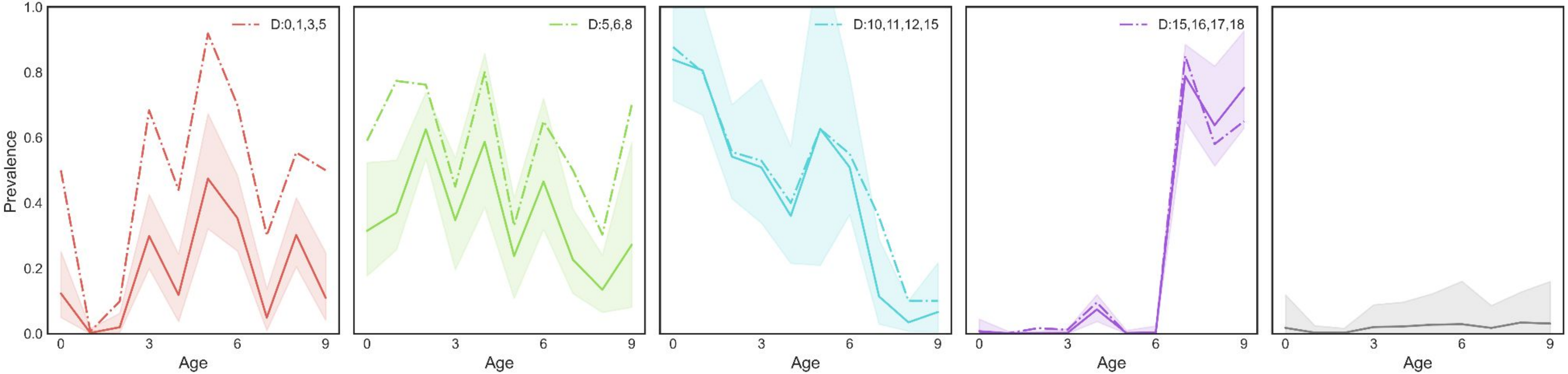
- Is model robust to mis-specified K?

- Experiments
  - Simulate the data from generative process.
  - Carry out posterior inference assuming that K is larger than its true counterpart.

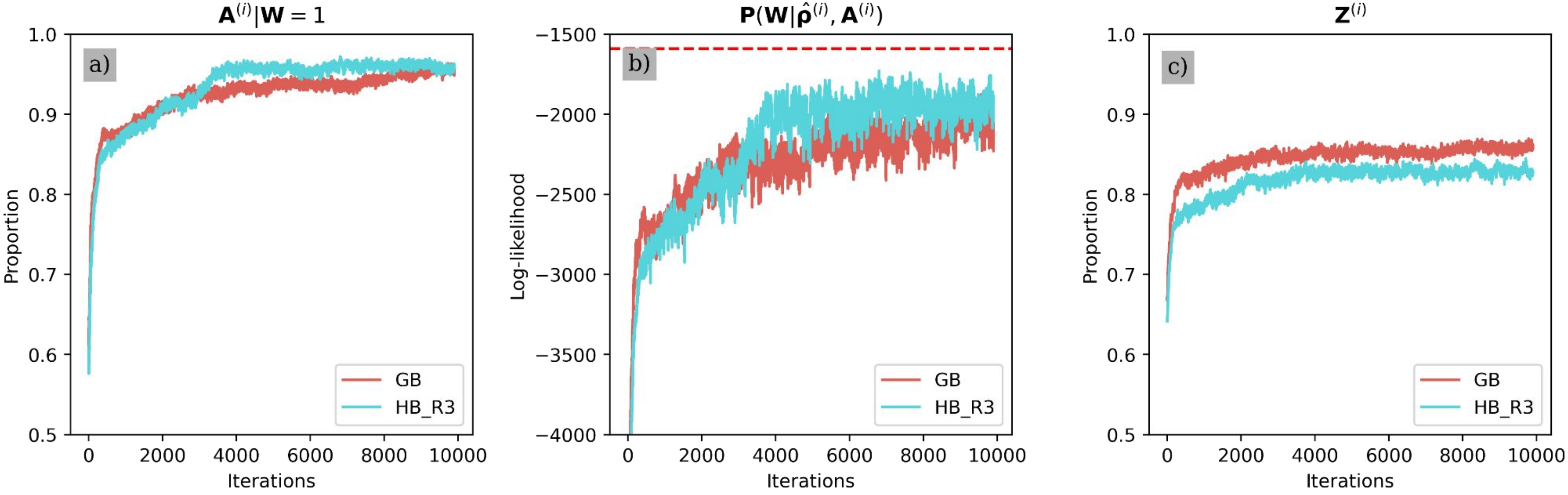
Posterior morbidity profiles



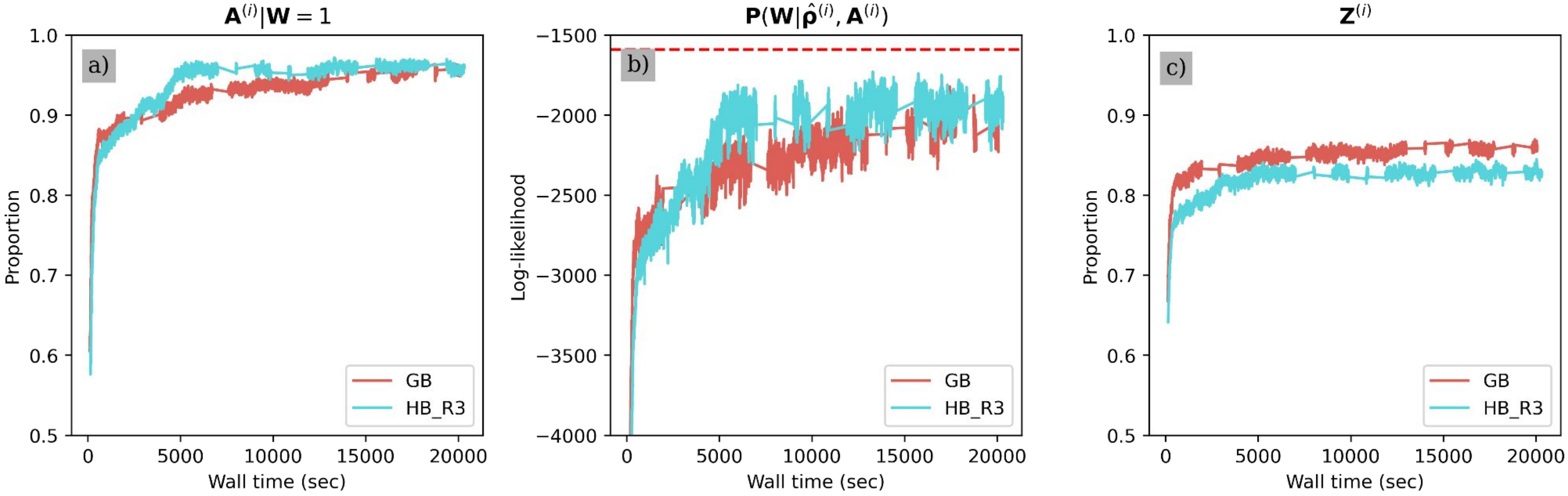
Posterior Multimorbidity Trajectory



HB vs GB (per iterations)



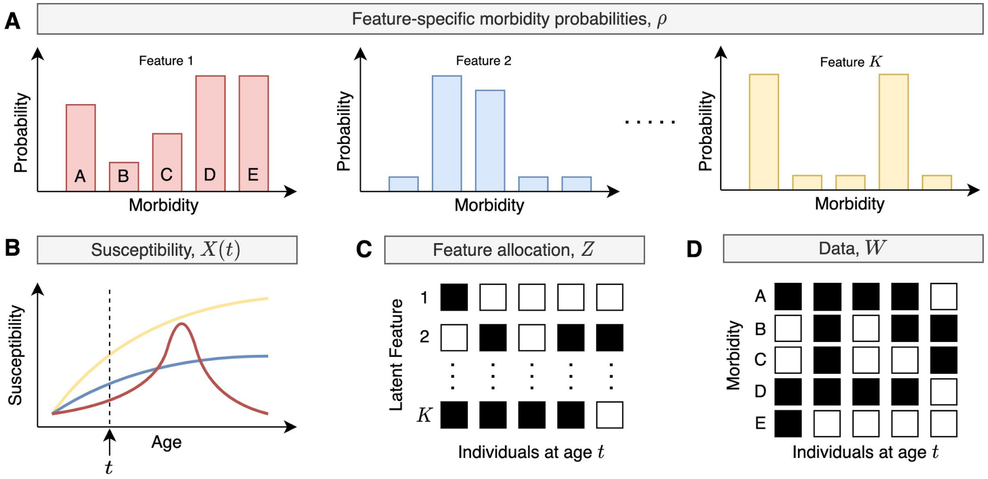
HB vs GB (per wall-clock)



Conclusions

What we have done:

1. Constructed a time-varying latent feature allocation model based on W-F diffusion process,
2. Used to model age-dependence in multimorbidity clustering,
3. Recapitulated patterns of age-linked dependence that we would expect to observe.



Further work

- Even more scalability with larger data sets (e.g. primary care data with millions of individuals).
- What to do with covariates (e.g., other demographic information, survival outcome, etc) ?

Multimodality Multimorbidity Variational Autoencoder

Woojung Kim
University of Warwick

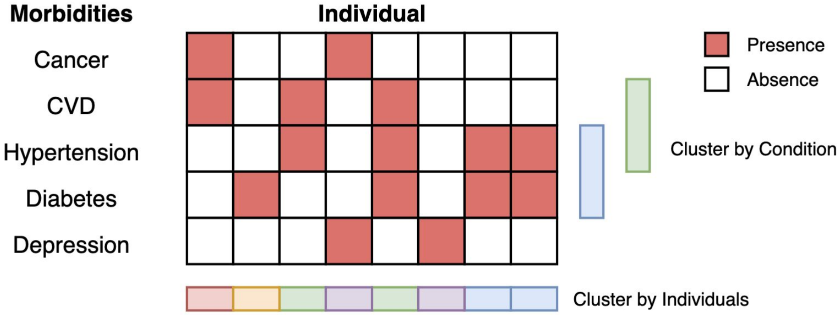
(with Paul Jenkins, University of Warwick and Christopher Yau, University of Oxford)

Existing works

Q. What are the conditions with the most significant health impacts?

Many studies have utilised dataset with a single modality (i.e. binary variables) to identify clusters of patients with similar morbities.

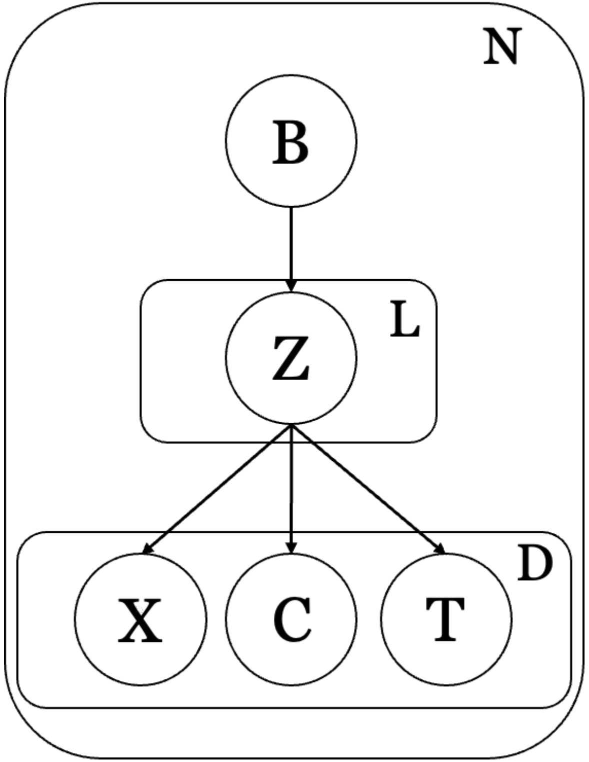
This constraint restricts the model's ability to gather information from multiple data sources (e.g. survival data, health measurements) to identify comorbidities with serious health impacts



Morbidities	Individual							
Cancer	🟥	□	□	🟥	□	□	□	□
CVD	🟥	□	🟥	□	🟥	□	□	□
Hypertension	□	□	🟥	□	🟥	□	🟥	🟥
Diabetes	□	🟥	□	□	🟥	□	🟥	🟥
Depression	□	□	□	🟥	□	🟥	□	□

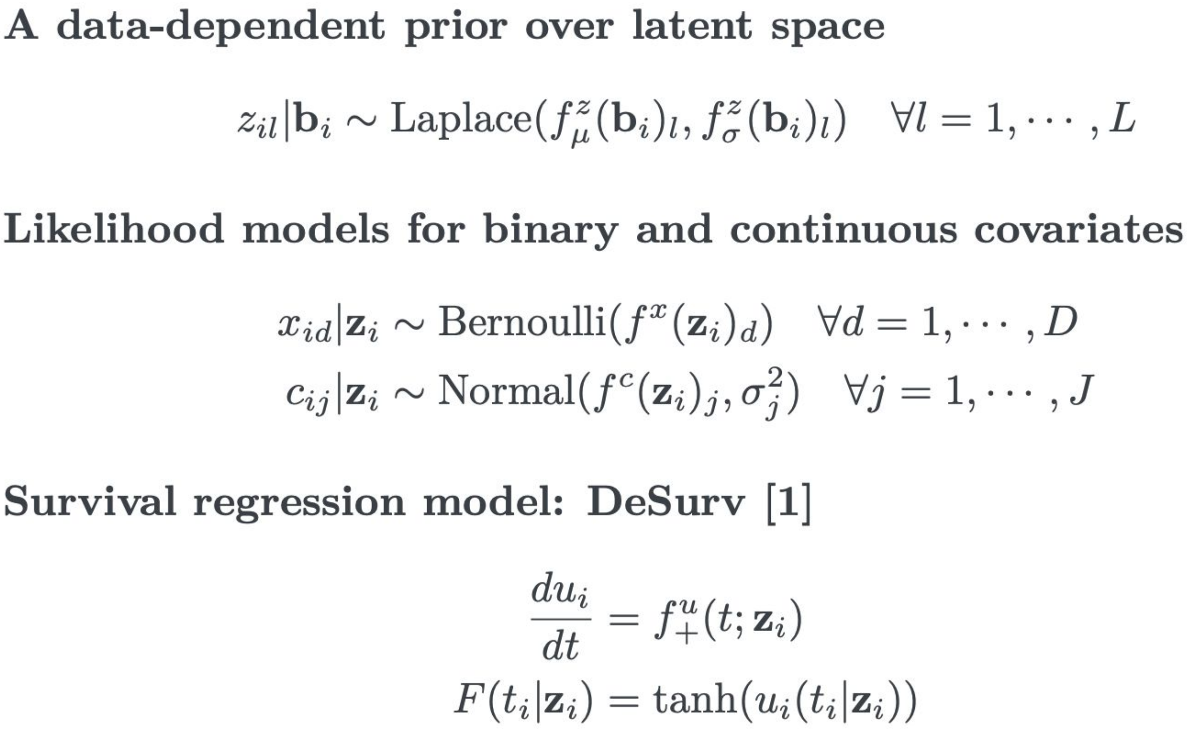
Multimodality Multimorbidity Variational Autoencoder

- B_{ib} - mixed-type demographic data
- z_{il} - continuous (low-dim) latent variables
- $X_{id} \in \{0, 1\}$ - binary morbidity data
- C_{ij} - continuous health measurements
- t_i - failure time



Goal is to obtain latent health summaries (Z) from mixed-type attributes (X,C,T) s.t. L < D

Generative process



A data-dependent prior over latent space

$$z_{il}|\mathbf{b}_i \sim \mathrm{Laplace}(f_\mu^z(\mathbf{b}_i)_l, f_\sigma^z(\mathbf{b}_i)_l) \quad \forall l = 1, \cdots, L$$

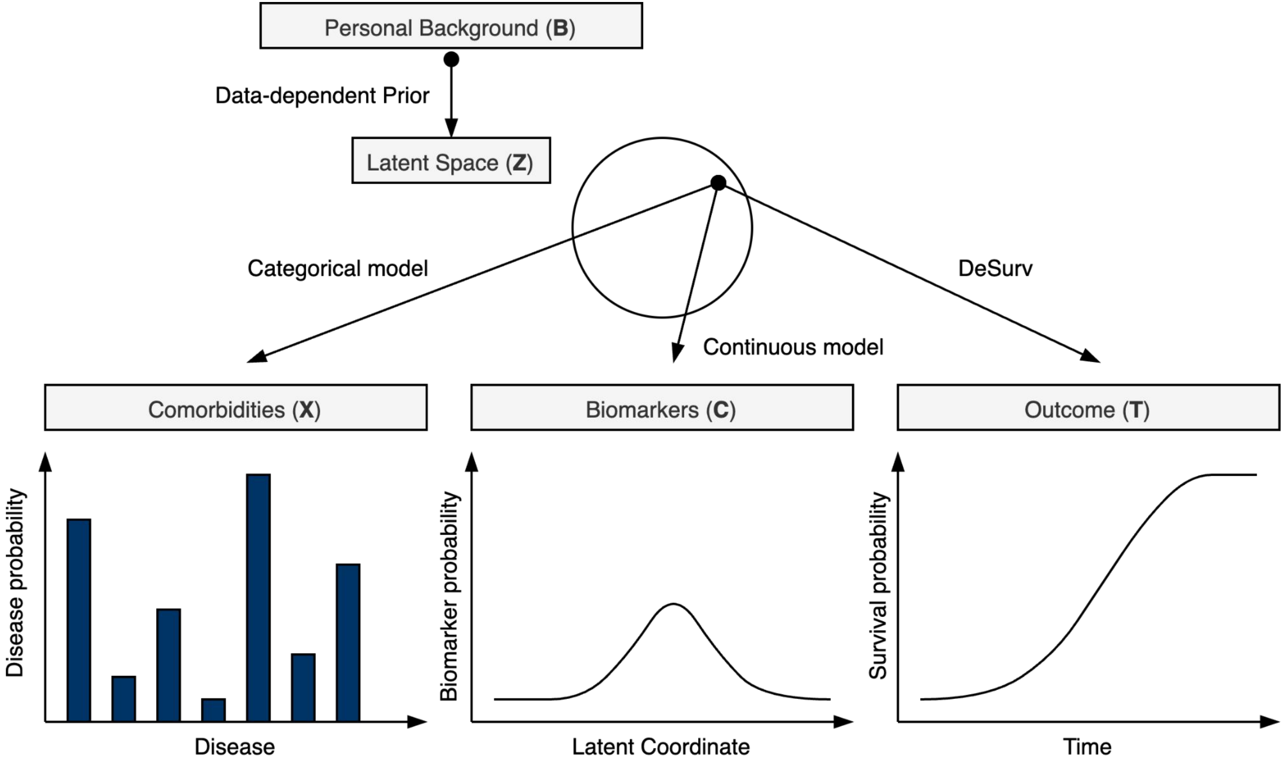
Likelihood models for binary and continuous covariates

$$x_{id}|\mathbf{z}_i \sim \mathrm{Bernoulli}(f^x(\mathbf{z}_i)_d) \quad \forall d = 1, \cdots, D$$
$$c_{ij}|\mathbf{z}_i \sim \mathrm{Normal}(f^c(\mathbf{z}_i)_j, \sigma_j^2) \quad \forall j = 1, \cdots, J$$

Survival regression model: DeSurv [1]

$$\frac{du_i}{dt} = f_+^u(t; \mathbf{z}_i)$$
$$F(t_i|\mathbf{z}_i) = \tanh(u_i(t_i|\mathbf{z}_i))$$

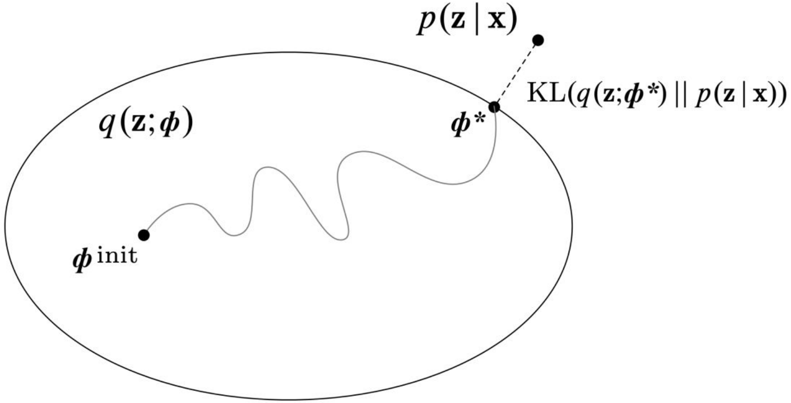
Generative Model



Rest of the analysis

- **Posterior inference is carried out using variational inference**
  - Choose an approximating distribution parameterized by (variational) parameters
  - Maximize a lower bound of the marginal data distribution with respect to both model and “variational” parameters
- **Analysis**
  - Evaluate the usefulness of latent representations

Variational Inference


$p(\mathbf{z} \mid \mathbf{x})$
$\mathrm{KL}(q(\mathbf{z};\boldsymbol{\phi}^*) \,||\, p(\mathbf{z} \mid \mathbf{x}))$
$q(\mathbf{z};\boldsymbol{\phi})$
$\boldsymbol{\phi}^*$
$\boldsymbol{\phi}^{\text{init}}$

- Construct a family of (tractable) probability distributions over latent variable z

$$Q = \{q_\phi(z) : \phi \in \Theta\}$$

- Optimise (variational) parameters to be close to the exact posterior (in KL divergence)

Variational inference solves inference with optimization

Evidence Lower Bound (ELBO)

Minimise

Posterior distribution

$$\mathrm{KL}(q_\phi(z)||p(z|x)) = \log p(x) - \underbrace{(\mathbb{E}_{z\sim q_\phi(z)}[\log p(x|z)] - \mathrm{KL}(q(z)||p(z)))}_{\textbf{ELBO}}$$

Variational distribution

Intractable!

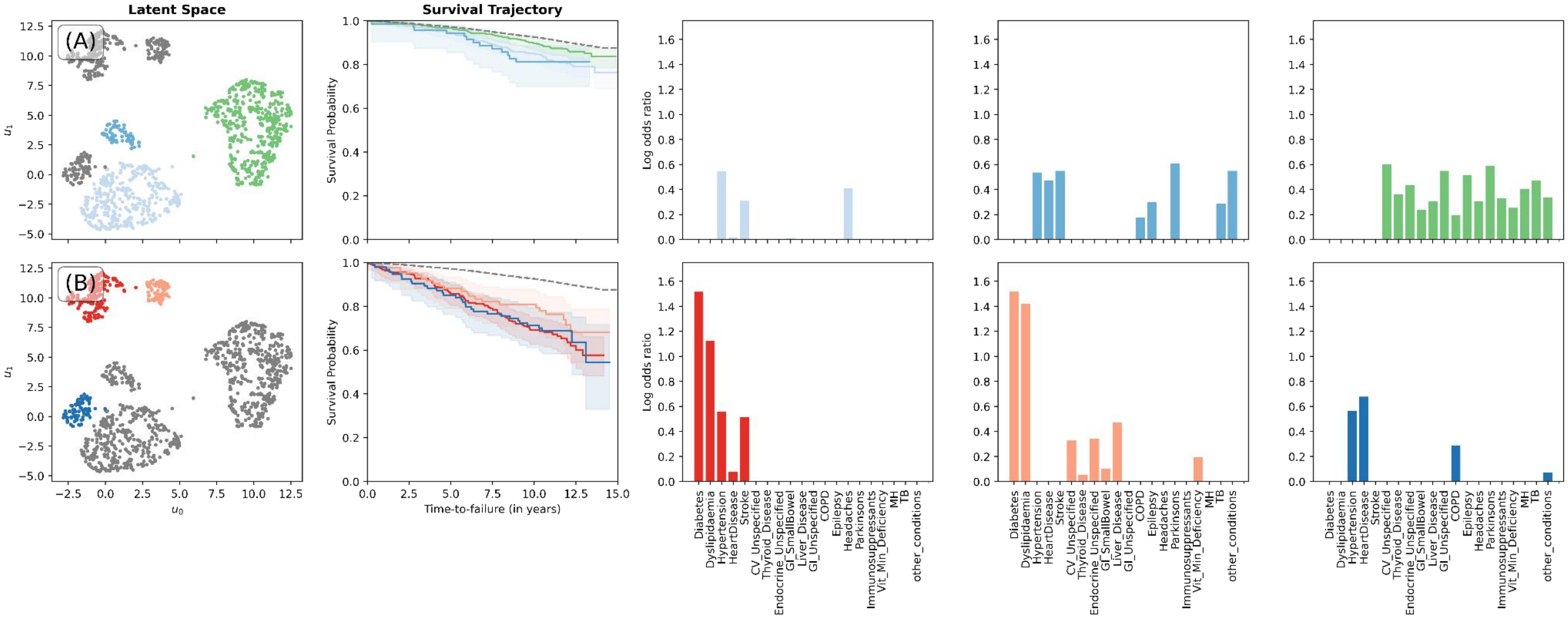
Maximize

The goal is to find a set of variational parameter ϕ to maximize the ELBO!

Usefulness of latent representations

1. Train the model.
2. Obtain latent representations from test-set data.
3. Apply UMAP (Leland et al, 2018) to project it into 2-dim space.
4. Apply K-means algorithm

Outcome


Latent Space
Survival Trajectory
(A)
(B)
u_0
u_1
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0.0
2.5
5.0
7.5
10.0
12.5
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
15.0
Time-to-failure (in years)
Log odds ratio
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Diabetes
Dyslipidaemia
Hypertension
HeartDisease
Stroke
CV_Unspecified
Thyroid_Disease
Endocrine_Unspecified
GI_SmallBowel
Liver_Disease
GI_Unspecified
COPD
Epilepsy
Headaches
Parkinsons
Immunosuppressants
Vit_Min_Deficiency
MH
TB
other_conditions

Summary

- Multimorbidity analysis aims to identify patterns of co-occurring morbidities, defined as:
    - Clusters of morbidities
    - Clusters of individuals
- Latent variable approaches can be a useful tool to identify both.

Q&A