M^{2}Traj: A Data-Driven Framework for Multimorbidity Trajectory Identification from Multi-Disease Diagnosis Time Series

Anonymous Author(s)

Abstract

Multimorbidity, the co-occurrence of multiple chronic conditions in an individual, is a growing global health challenge, particularly in ageing populations. Although existing studies have identified cross-sectional patterns of multimorbidity, little is known about how these patterns evolve throughout life. In this study, we propose M^{2}Traj, a data-driven framework to identify and analyse multimorbidity trajectories from large-scale longitudinal diagnosis records. Leveraging electronic health records from 3.3 million individuals in England, we first stratify participants by age and sex, then apply latent class analysis to learn disease co-occurrence clusters. We merge similar clusters across age bands using hierarchical clustering to define stable multimorbidity profiles, and reconstruct individual-level trajectories by tracking transitions between profiles over time. Our results reveal interpretable progression pathways, persistent profiles, and critical transition points across the lifespan, providing novel insights to support long-term prevention, risk stratification, and chronic disease management.

Keywords

Event Time Series, Clustering, Latent Class Analysis, Healthcare

ACM Reference Format:

Anonymous Author(s). 2018. M^{2}Traj: A Data-Driven Framework for Multimorbidity Trajectory Identification from Multi-Disease Diagnosis Time Series. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The digitalisation of healthcare systems has led to the accumulation of large-scale electronic health records (EHRs), providing unprecedented opportunities for artificial intelligence (AI) and data mining to transform healthcare research and clinical decision-making [13]. In particular, longitudinal EHR data offer a rich resource for identifying patient subgroups, modelling disease progression, and supporting personalised and population-level care strategies.

One of the most critical challenges in this context is multimorbidity, which refers to the presence of two or more chronic conditions in an individual [15, 16]. Multimorbidity is highly prevalent, particularly in ageing populations, and is associated with poorer health outcomes, greater use of healthcare, and increased costs [16]. Although many studies quantify multimorbidity simply by counting co-occurring conditions [6, 7, 9], such measures do not account for heterogeneity in disease composition and progression. To address this, recent research has focused on identifying multimorbidity profiles, i.e., common combinations of chronic conditions. These studies have discovered clinically significant disease groups such as cardiovascular or cardiometabolic profiles [2–4, 14]. However, most of them are based on cross-sectional data and cannot capture how disease patterns evolve over the life course.

Studying life-course multimorbidity trajectories presents unique methodological challenges. First, diagnosis data are sparse and irregularly sampled, with high dimensionality and variable observation windows. Furthermore, disease development is influenced by age and sex, requiring stratified modelling. To address these challenges, we propose a scalable, interpretable framework named M^{2}Traj for identifying and reconstructing multimorbidity trajectories from multi-disease diagnosis time series data. Specifically, M^{2}Traj first converts multi-disease diagnosis histories into age-based representations and models the health states of individuals as binary vectors indicating the presence or absence of multiple conditions. Within each age-sex group, a latent class analysis [1, 5, 11] is applied to discover clusters of individuals who share similar disease patterns. These clusters are then linked across adjacent age bands using hierarchical clustering based on disease prevalence and exclusivity to define multimorbidity profiles that are stable over time. Finally, individual trajectories are reconstructed by tracking transitions between profiles across the life span, enabling downstream analyses of disease progression pathways at the population scale.

Our key contributions are summarised as the following:

- We propose M^{2}Traj, a scalable and interpretable framework for multimorbidity profile and trajectory identification from multi-disease diagnosis time series.
- We apply latent class analysis and hierarchical clustering to identify 39 multimorbidity profiles stratified by age and sex, capturing distinct and clinically meaningful combinations of chronic conditions in primary care.
- We reconstruct population-wide multimorbidity trajectories by tracing individual-level transitions between profiles over nine age bands, revealing dominant pathways and critical stages of disease accumulation.

2 Dataset Description

We used primary care electronic health records on a population scale in England to examine the distribution and progression of multimorbidity. Specifically, we accessed data from the UK Clinical Practice Research Datalink (CPRD) [18], identifying 3,314,652 individuals who had been diagnosed with two or more chronic conditions by the end of 2019. For each individual, longitudinal diagnosis records were available from the time of registration in

Conference acronym 'XX,

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

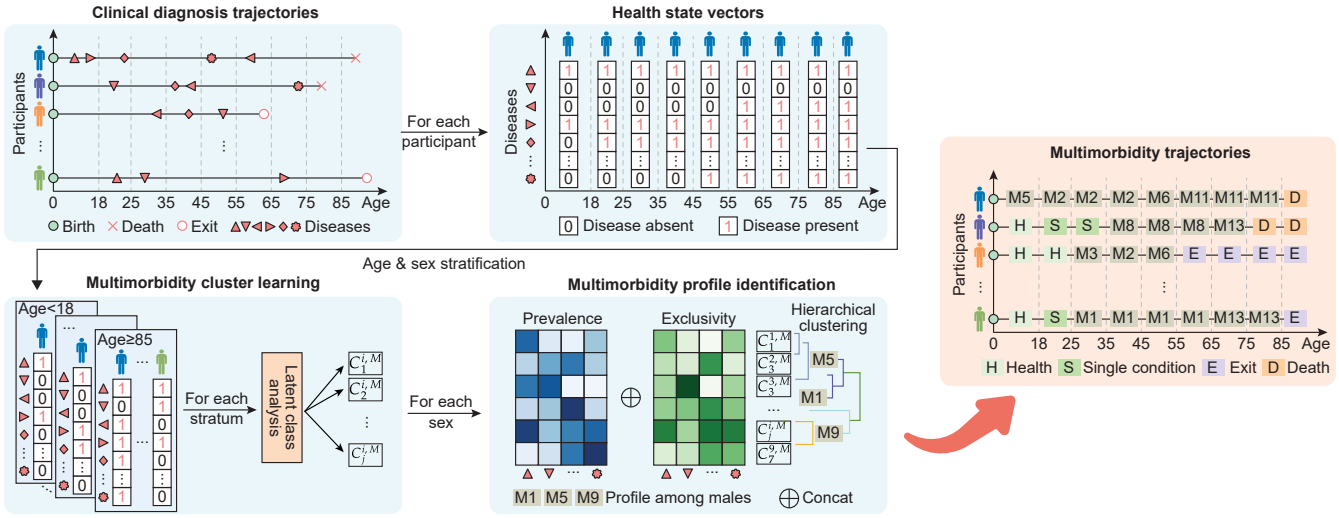

Figure 1: Illustration of our proposed M^{2}Traj framework, shown for male patients. Clinical diagnosis trajectories are extracted for participants over the life course, stratified by sex and age band. Trajectories are encoded as health state vectors representing the presence of specific conditions at each age band. Latent class analysis is applied within each stratum to identify multimorbidity clusters (e.g., $C_j^{i,M}$ denotes the j-th cluster within the i-th age band for males). To capture consistent multimorbidity patterns across age bands, hierarchical clustering is performed based on condition prevalence and exclusivity within the cluster. Condition prevalence is defined as the proportion of individuals within a cluster who has a given condition, whereas condition exclusivity is defined as the proportion of individuals with a specific condition in a given age band who belongs to that particular cluster. This yields generic multimorbidity profiles (e.g., M1, M5, M9). Individual multimorbidity trajectories are reconstructed by mapping transitions across profiles over the life course.

the healthcare system, from birth until death or exit from the study. Study exit was defined as the first deregistration from general practice or the last data collection date for that practice. Our dataset includes longitudinal diagnoses of 18 commonly reported chronic diseases spanning mental health (anxiety, depression, serious mental illnesses), respiratory (asthma, chronic obstructive pulmonary disease), metabolic (diabetes), cardiovascular (hypertension, coronary heart disease, stroke or transient ischaemic attack, atrial fibrillation, heart failure, peripheral arterial disease), renal (chronic kidney disease), neurological (dementia, Parkinson's disease), musculoskeletal (osteoporosis, rheumatoid arthritis), and oncological (cancers) domains.

3 Methods

3.1 Problem Definition

We are given a patient diagnosis dataset $\mathcal{D} = \{(i, \mathcal{D}_i) \mid i = 1, \cdots, N\}$, where N is the number of patients. The multi-disease diagnosis history of the i-th patient is defined as an event time series (i.e., irregularly sampled time series) $\mathcal{D}_i = \{(x_{i,j}, t_{i,j}) \mid j = 1, \cdots, N_i^{\text{disease}}\}$, where $x_{i,j}$ and $t_{i,j}$ denote the type of disease and the diagnosis time of the j-th disease diagnosed with the patient i. The type of disease $x_{i,j}$ belongs to a set of N^{disease} unique diseases $\mathcal{X}$. Hence, the multimorbidity trajectory task can be formally defined below.

Problem 1. **Multimorbidity Trajectory Identification.** *Given a diagnosis dataset $\mathcal{D}$, the multimorbidity trajectory identification task is to identify the progression pathways across a set of profiles $\mathcal{P} = \{P_j \mid j = 1, \cdots, N^{\text{profile}}\}$ throughout the life course, such that each profile P_j represents a group of patients with similar patterns of chronic diseases.*

An overview of the proposed framework, M^{2}Traj, is shown in Figure 1. For each individual, we first transform their diagnosis history into an age-aligned trajectory and construct a sequence of health state vectors across predefined age bands. Within each age band, latent class analysis is used to group individuals into disease clusters. To ensure longitudinal consistency, we compute similarity metrics across clusters of adjacent age bands and apply hierarchical clustering to derive a unified set of multimorbidity profiles. Finally, each individual's trajectory is reconstructed by mapping their progression across these profiles over time.

3.2 Stratified Diagnosis Representations

Age and sex are known to influence the development of multimorbidity [10, 15]. We therefore stratify our analysis by age and sex. For each individual i, the diagnosis trajectory $\mathcal{D}_i = \{(x_{i,j}, t_{i,j})\}$ is converted into an age-based trajectory $\mathcal{D}_i^{\text{age}} = \{(x_{i,j}, a_{i,j})\}$, where $a_{i,j}$ is the age at diagnosis. The lifespan is divided into nine age bands: <18, 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and ≥85, following previous research [19]. For each band a and sex $s \in \{M, F\}$, an individual's health state is represented as a binary vector $\boldsymbol{h}_i^{a,s} \in \{0,1\}^{N^{\text{disease}}}$ indicating the presence or absence of each disease.

3.3 Multimorbidity Cluster Learning

For each age-sex stratum (a, s), we develop a latent class analysis method [11] for clustering. Specifically, let $\mathcal{H}^{a,s} = \{h_i^{a,s} \mid i = 1, \cdots, N^{a,s}\}$ denote the set of health state vectors of all the $N^{a,s}$ patients observed in that stratum. Given a predetermined number of latent classes (clusters) K, each health state vector $\boldsymbol{h}_i^{a,s}$ is assumed to be generated conditionally independently given the latent class $C \in \{1, \cdots, K\}$, which can be expressed as:

$$\mathbb{P}(\boldsymbol{h}_i^{a,s} \mid C = k) = \prod_{j=1}^{N^{\text{disease}}} \mathbb{P}(h_{i,j}^{a,s} \mid C = k).$$

Furthermore, the likelihood of the observation is:

$$\mathbb{P}(\boldsymbol{h}_i^{a,s}) = \sum_{k=1}^{K} \mathbb{P}(C = k)\mathbb{P}(\boldsymbol{h}_i^{a,s} \mid C = k).$$

Let $\theta_{j,k} = \mathbb{P}(h_j = 1 \mid C = k)$ denote the class-specific Bernoulli probability for disease j, and let $\pi_k = \mathbb{P}(C = k)$ be the prior probability of class k, satisfying $\sum_{k=1}^{K} \pi_k = 1$. For simplicity, we omit the stratum identifier (a, s) for the two variables. Then the marginal likelihood for an individual $\boldsymbol{h}_i^{a,s}$ is expressed as:

$$\mathbb{P}(\boldsymbol{h}_i^{a,s}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{N^{\text{disease}}} \theta_{jk}^{h_{i,j}^{a,s}} (1 - \theta_{j,k})^{1 - h_{i,j}^{a,s}}.$$

The total log-likelihood over all patients in stratum (a, s) is then calculated as:

$$\mathcal{L}^{a,s} = \sum_{i=1}^{N^{a,s}} \log \left(\sum_{k=1}^{K} \pi_k \prod_{j=1}^{N^{\text{disease}}} \theta_{jk}^{h_{i,j}^{a,s}} (1 - \theta_{j,k})^{1 - h_{i,j}^{a,s}} \right).$$

The model parameters $\{\pi_k, \theta_{j,k} \mid k = 1, \cdots, K, j = 1, \cdots, N^{\text{disease}}\}$ are estimated by maximising this log-likelihood using the Expectation-Maximization (EM) algorithm. Finally, we derive the posterior probability that an observation $\boldsymbol{h}_i^{a,s}$ belongs to latent class k as:

$$\gamma_{i,k}^{a,s} = \mathbb{P}(C = k \mid \boldsymbol{h}_i^{a,s}) = \frac{\pi_k \prod_{j=1}^{N^{\text{disease}}} \theta_{j,k}^{h_{i,j}^{a,s}} (1 - \theta_{j,k})^{1 - h_{i,j}^{a,s}}}{\sum_{l=1}^{K} \pi_l \prod_{j=1}^{N^{\text{disease}}} \theta_{j,l}^{h_{i,j}^{a,s}} (1 - \theta_{j,l})^{1 - h_{i,j}^{a,s}}}.$$

Therefore, the most likely cluster assignment for each patient is $\arg\max_k \gamma_{i,k}^{a,s}$, yielding K clusters per stratum. We let $C_j^{a,s}$ denote the j-th obtained cluster from age band a and sex group s.

Following previous studies [1, 5], the optimal number of clusters for each stratum is determined based on model parsimony using the Bayesian information criterion (BIC), Akaike information criterion (AIC) and consistent AIC (cAIC), which balance goodness of fit against model complexity to minimise overfitting. The final selection also incorporates clinical relevance and interpretability, as established through successive rounds of review by an expert panel and consensus meetings with clinicians.

3.4 Multimorbidity Profile Identification

Once multimorbidity clusters have been identified within each age-sex stratum (a, s), we examine their consistency across adjacent age bands to identify stable patterns of disease accumulation. Since chronic diseases often exhibit persistent trends with ageing, clusters from neighbouring age bands may represent similar underlying multimorbidity patterns. To capture these longitudinal consistencies, we implement a cluster merging step to construct the final set of multimorbidity profiles.

We merge clusters based on two key characteristics: disease prevalence and disease exclusivity. For a given cluster $C_j^{a,s}$, let $\mathcal{H}_j^{a,s} = \{\boldsymbol{h}_i^{a,s} \mid \arg\max_k \gamma_{i,k}^{a,s} = j\}$ denote the set of health state vectors assigned to that cluster. The disease prevalence vector $\boldsymbol{f}_j^{a,s}$ represents the proportion of individuals within the cluster diagnosed with each disease and is defined as:

$$\boldsymbol{f}_j^{a,s} = \frac{1}{|\mathcal{H}_j^{a,s}|} \sum_{\boldsymbol{h} \in \mathcal{H}_j^{a,s}} \boldsymbol{h}.$$

On the other hand, the disease exclusivity vector is calculated as the proportion of individuals with a specific disease in a given age band who belong to that particular cluster, and the exclusivity value for l-th disease can be calculated as:

$$\boldsymbol{g}_{j,l}^{a,s} = \frac{\sum_{\boldsymbol{h} \in \mathcal{H}_j^{a,s}} h_l}{\sum_{\boldsymbol{h} \in \mathcal{H}^{a,s}} h_l}.$$

With the two characteristic vectors obtained, we concatenate them together into a vector, i.e., $\boldsymbol{x}_j^{a,s} = [\boldsymbol{f}_j^{a,s}; \boldsymbol{g}_j^{a,s}] \in \mathbb{R}^{2N^{\text{disease}}}$, which describes the characteristics of the j-th multimorbidity cluster learnt from the age-sex stratum (a, s). Subsequently, we apply agglomerative hierarchical clustering using Ward's method to the concatenated vectors of each sex group, thus quantifying the similarity of the cluster across age bands. Clusters within each sex group are merged based on thresholds informed by clinical interpretability and hierarchical clustering results, thus forming multimorbidity profiles $\mathcal{P} = \{P_i \mid 1, \cdots, N^{\text{profile}}\}$. Each multimorbidity profile is characterised by its distinct composition of chronic diseases and is named according to a convention guided by clinicians. This approach ensures accurate identification of sex- and age-specific profiles while capturing multimorbidity patterns consistently across different strata.

3.5 Multimorbidity Trajectory Reconstruction

Based on the multimorbidity profiles identified in each age-sex stratum, we reconstruct the temporal progression of each individual as a discrete-time sequence over predefined age bands. Specifically, for each individual, we construct a sequence of states, where each state corresponds to one of the following:

- a multimorbidity profile (e.g., $P_1, P_2, \ldots, P_{N^{\text{profile}}}$),
- a healthy state without any of the 18 studied conditions (denoted as "H")
- a state with only one diagnosed condition (denoted as "S")
- death (denoted as "D")
- study exit (denoted as "E").

This results in a trajectory vector of length equal to the number of age bands (nine in our study), capturing the patient's longitudinal evolution of the patient's health states throughout life. The structure of these trajectories is discrete, aligned to irregular age-specific sampling windows, and categorical in nature, making it particularly suitable for symbolic time series analysis.

To enable downstream analysis, we further encode each trajectory as a symbolic sequence $T_i = [z_i^1, z_i^2, ..., z_i^A]$, where $z_i^a \in \mathcal{Z}$ is the assigned state of the individual i in the age band a, and $\mathcal{Z}$ is the full set of possible states as described above. This symbolic representation allows for:

- **Trajectory pattern discovery**, using frequent sequence mining techniques to identify common progression paths;
- **Temporal clustering**, where patients are grouped based on shared progression patterns using distance measures such as Dynamic Time Warping (DTW) or edit distance;
- **Predictive modelling**, enabling future state prediction (e.g., transition to a complex multimorbidity profile) using Markov models or sequence-based neural architectures;
- **State transition analysis**, where we estimate empirical transition matrices and characterise dominant disease accumulation pathways across age bands.

Importantly, this formulation generalises traditional time series analysis to the setting of discrete multivariate event sequences with age-aligned structure. It accommodates the irregular nature of diagnosis data and uses unsupervised temporal abstraction to model progression without requiring dense sampling or fixed-length time intervals. Hence, M^{2}Traj provides a scalable and interpretable framework for mapping population-wide multimorbidity dynamics using real-world multi-disease diagnosis time series.

4 Experiments

4.1 Experimental Setting

We use the dataset described before for experiments. The prevalence of the multimorbidity profile within the social subgroup is used to quantify social disparities. All experiments are run on an RTX 6000 GPU with 32 GB RAM. The latent class analysis methods are implemented using the StepMix package [12], and hierarchical clustering is performed using the SciPy package [17] in Python.

4.2 Identified Multimorbidity Profiles

The proposed framework identified 21 male and 18 female multimorbidity profiles over the course of life, as shown in Figure 2. These profiles exhibited a clear age-dependent progression, increasing in both prevalence and complexity with age.

In early life (<35 years), profiles were primarily mental health–predominated, often co-occurring with asthma (e.g., M1–M5 in males, F1–F6 in females). From age 35 to 64, the cardiometabolic and cancer-related profiles became more prominent (M6–M12, F7–F13), while older adults (≥65) exhibited more complex, multi-system profiles that involve cardiovascular, renal, metabolic, and respiratory domains (M13–M21, F14–F18).

Profiles classified as "complex" (mean condition count >4 [8]) were concentrated in older age bands. For example, M14 (Cardiovascular + Cardiometabolic + Renal) and M15 (Mental + Cardiovascular + Renal + Respiratory) in males had mean condition counts of 5.86 (SD: 1.34) and 7.03 (SD: 1.30), respectively. Their female counterparts, F16 and F15, exhibited similar complexity, with 4.53 and 7.02 conditions on average. Distinct disease combinations were evident: while M15 and F15 were dominated by mental health and respiratory conditions (depression prevalence >85%), M14 and F16 had a higher burden of heart failure and atrial fibrillation.

Several profiles showed long persistence across age bands. M1 and F1 (Anxiety + Depression) were present across six age bands before 65. F10 (Cardiometabolic + Renal) and M11 (CHD-predominant Cardiovascular + Diabetes) spanned five consecutive bands starting at age 45. F10 ultimately became the most prevalent profile in older females, peaking at 28.24% (95% CI: 28.14–28.35%) in ages 75–84.

Sex-specific patterns were consistent and notable. Males were overrepresented in cardiovascular, respiratory, and cancer-dominant profiles (e.g., M11, M14–M16), while females appeared more frequently in mental, neurological, and musculoskeletal profiles (e.g., F14, F15, F18). For example, M19 (predominant in COPD) had a prevalence of COPD of 81. 37%, compared to a lower respiratory complexity in the female profiles. In contrast, osteoporosis was concentrated in F13 (Musculoskeletal + Cancer), with a prevalence exceeding 94%.

These findings highlight consistent temporal patterns in multimorbidity evolution and marked differences by sex and disease domain. The profiles offer a data-driven abstraction of disease co-occurrence, which serve as the foundation for reconstructing trajectories in the next stage.

4.3 Identified Multimorbidity Trajectories

Figure 3 presents the reconstructed multimorbidity trajectories across the age bands using the symbolic representations derived from the profile assignments. The number of individuals in multimorbidity states increased with age, peaking at 65–74 and declining in older bands due to mortality or exit. The transitions between profiles reveal dominant pathways of disease progression throughout life.

In both sexes, mental health-only profiles (M1 and F1) frequently marked the onset of multimorbidity and served as precursors to more complex profiles that incorporate hypertension and diabetes. By midlife (45–64), transitions from M1/F1 to M6/F7 (Hypertension + Depression + Anxiety) became common. In later age bands (65–74), profiles such as M13 and F14 (Mental + Physical Long-term Conditions) absorbed individuals from M1, F1, M6, F7, and M8/F8, indicating convergence toward multi-system burden.

Cardiovascular trajectories typically involved early transitions from healthy or single-disease states into M11 or F11 (CHD + Diabetes ± Mental Health), followed by progression to more complex cardiometabolic-renal profiles such as M14 and F16. For example, M11 contributed 42.25% and 31.22% of M14's population in the 65–74 and 75–84 bands, respectively. A similar sequence was observed in females: F11 transitioned into F16 through F10 and F17 intermediaries.

Cardiometabolic-renal transitions exhibited shared and sex-specific patterns. In males, M10 (Hypertension + Diabetes) and M12 (Hypertension + Cancer + Renal) were frequent entry points, often leading to M17 (Cardiometabolic + Renal) and M18 (Cancer + Physical Long-term Conditions). These two profiles accounted for over 70% of M17 and nearly 80% of M18. In females, F10 and F17 played analogous roles, feeding into F16, with 45.50% of F17's population transitioning to F10 in the 75–84 age band.

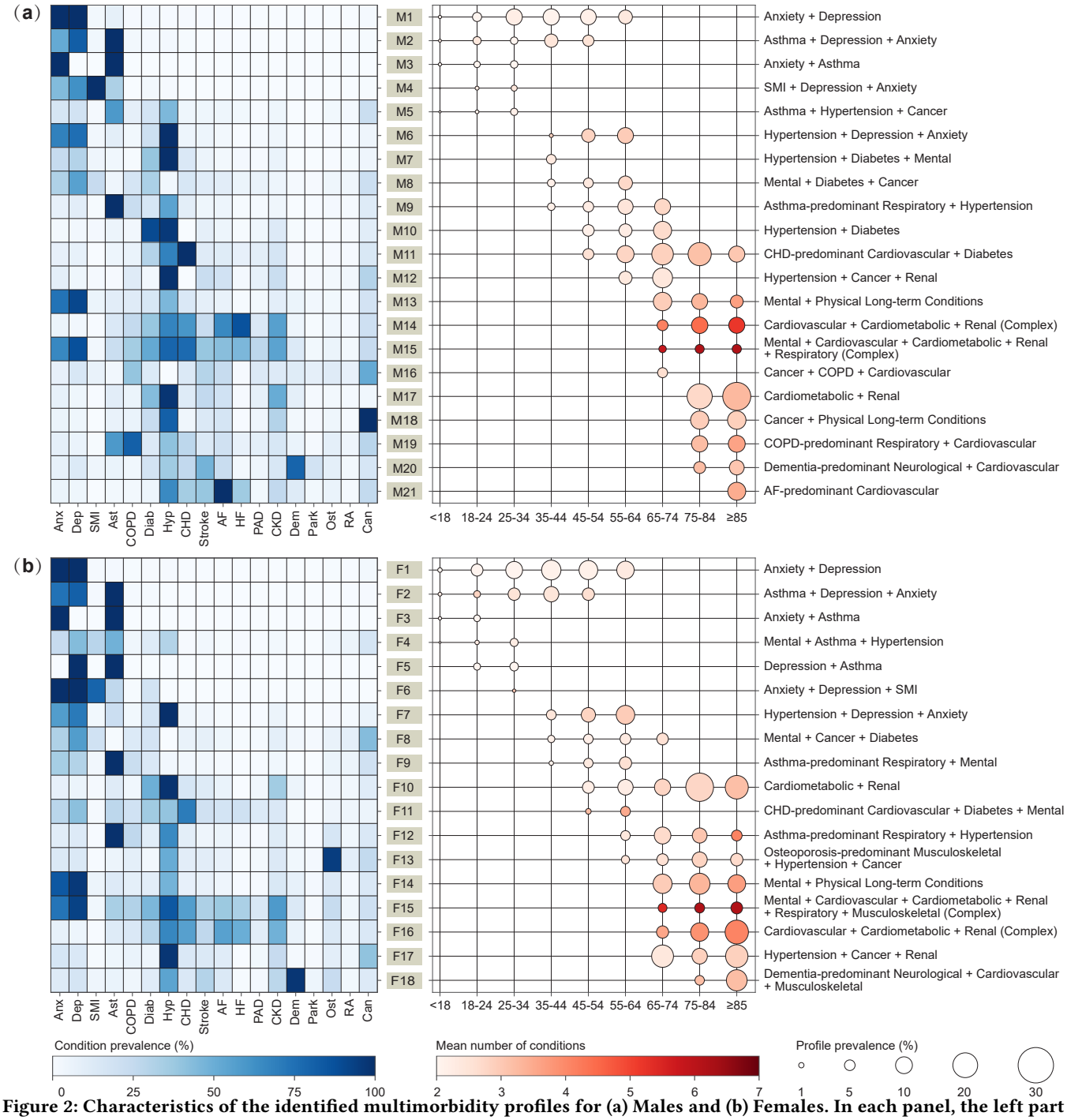

Figure 2: Characteristics of the identified multimorbidity profiles for (a) Males and (b) Females. In each panel, the left part presents a heatmap of disease prevalence within each profile, with darker blue shades indicating higher prevalence. The middle part presents a bubble plot, where the size of circles corresponds to the prevalence for a profile across an age band. The colour intensity reflects the mean number of conditions per individual, with darker red shades indicating a higher number of conditions. Profile labels are positioned between the two plots, and the right part lists the name of each profile. Diseases listed in the profile names are ordered by prevalence, and those with a mean number of diseases exceeding four are annotated as "Complex". Condition abbreviations: anxiety (Anx), depression (Dep), serious mental illness (SMI), asthma (Ast), chronic obstructive pulmonary disease (COPD), diabetes (Diab), hypertension (Hyp), coronary heart disease (CHD), stroke or transient ischaemic attack (Stroke), atrial fibrillation (AF), heart failure (HF), peripheral arterial disease (PAD), chronic kidney disease (CKD), osteoporosis (Ost), rheumatoid arthritis (RA), cancer excluding non-melanoma skin cancers (Can).

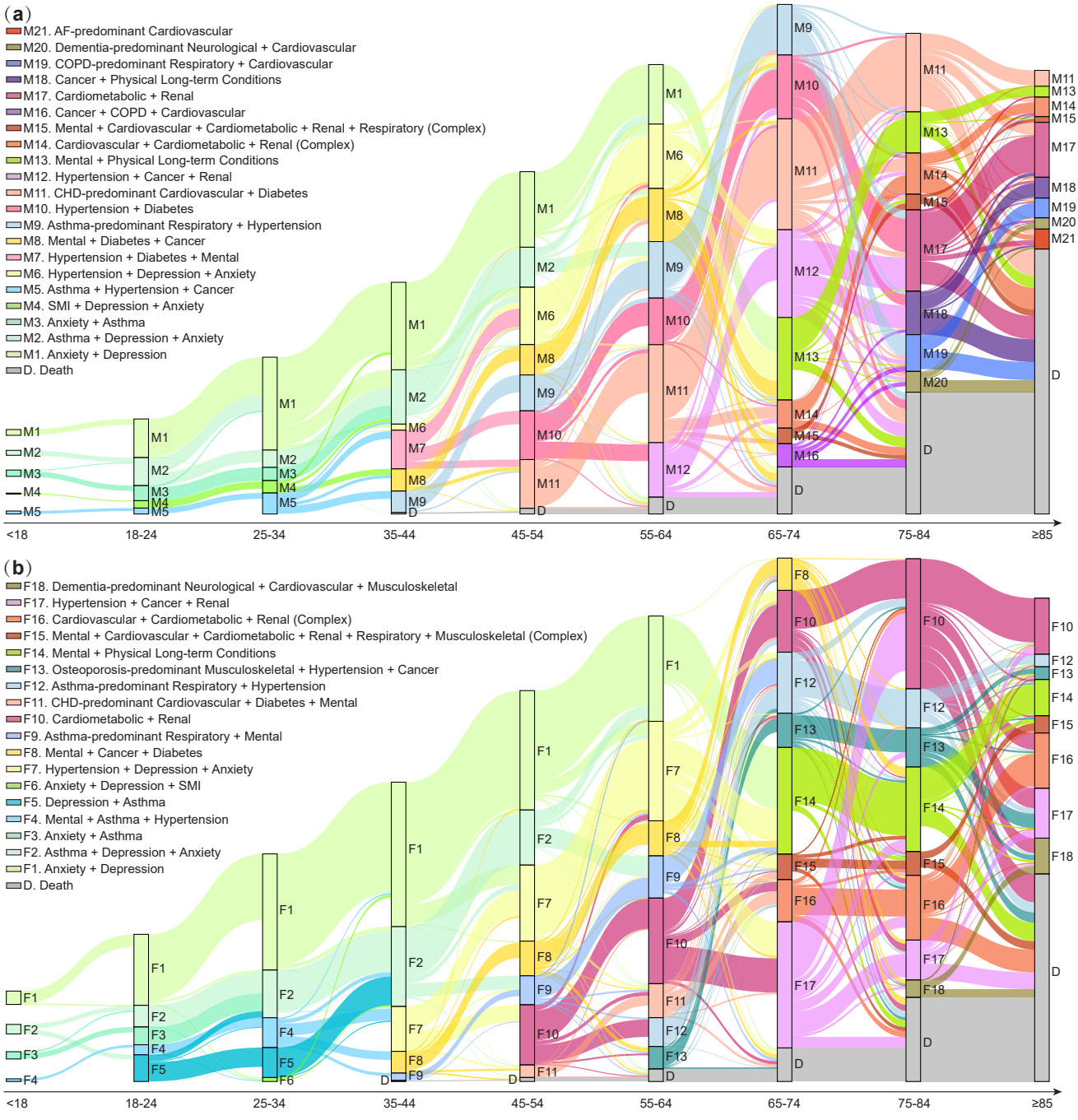

Figure 3: Multimorbidity trajectories over the life course for (a), Males and (b), Females. The Sankey diagrams illustrate transitions of the 3.3 million individuals in the primary study cohort between multimorbidity profiles across different age bands. For each panel, the height of each bin is proportional to the number of individuals within the corresponding profile. Each coloured flow represents the transition of individuals from one profile to another, where the thickness of the flow is proportional to the number of individuals in the transition. Transitions from the same source profile are shown in the same colour. The common profiles across sexes are represented using the same colour, and profiles with similar condition compositions use similar hues. For clarity, transitions from individuals with no or single conditions into multimorbidity profiles are omitted.

Respiratory trajectories also differed by sex. In males, M9 (Asthma + Hypertension) was present from ages 35–74 and frequently transitioned to M19 (COPD + Cardiovascular) at older age. In females, respiratory and mental health conditions were more closely coupled. F5 (Depression + Asthma) transitioned almost entirely to F2 (Anxiety + Depression + Asthma) in midlife, which then contributed heavily to F9 and F12, both asthma-predominant with increasing physical comorbidities.

Overall, the profile-to-profile transitions provide an interpretable discrete approximation of disease progression. Common patterns include: (1) mental health–only profiles as early-stage multimorbidity; (2) cardiometabolic conditions emerging in midlife and persisting into later stages; (3) distinct paths toward complexity through either cardiovascular or respiratory comorbidity accumulation. The identified multimorbidity trajectory captures critical transition points over the life course, offering a principled way to stratify the population by risk and temporal dynamics.

5 Conclusion

We present M^{2}Traj, a comprehensive framework for identifying multimorbidity profiles and reconstructing their longitudinal trajectories using large-scale primary care records of multi-disease diagnosis time series. By integrating latent class analysis, hierarchical clustering, and age–sex stratified trajectory modelling, our approach captures the dynamic evolution of multimorbidity across the life course. The resulting profiles reveal persistent disease clusters, age-related progression patterns, and clear sex-specific differences. These findings provide actionable insights for tailoring prevention and intervention strategies, underscoring the value of life-course-informed and population-scale approaches to chronic disease management.

References

[1] Javier Álvarez-Gálvez, Esther Ortega-Martín, Jesús Carretero-Bravo, Celia Pérez-Muñoz, Víctor Suárez-Lledó, and Begoña Ramos-Fiol. 2023. Social determinants of multimorbidity patterns: a systematic review. *Frontiers in Public Health* 11 (2023), 1081518.

[2] Javier Alvarez-Galvez, Esther Ortega-Martin, Begoña Ramos-Fiol, Victor Suarez-Lledo, and Jesus Carretero-Bravo. 2023. Epidemiology, mortality, and health service use of local-level multimorbidity patterns in South Spain. *Nature Communications* 14, 1 (2023), 7689.

[3] Ljoudmila Busija, Karen Lim, Cassandra Szoeke, Kerrie M Sanders, and Marita P McCabe. 2019. Do replicable profiles of multimorbidity exist? Systematic review and synthesis. *European Journal of Epidemiology* 34, 11 (2019), 1025–1053.

[4] Jonás Carmona-Pírez, Antonio Gimeno-Miguel, Kevin Bliek-Bueno, Beatriz Poblador-Plou, Jesús Díez-Manglano, Ignatios Ioakeim-Skoufa, Francisca González-Rubio, Antonio Poncel-Falcó, Alexandra Prados-Torres, Luis A Gimeno-Feliu, et al. 2022. Identifying multimorbidity profiles associated with COVID-19 severity in chronic patients using network analysis in the PRECOVID Study. *Scientific Reports* 12, 1 (2022), 2831.

[5] Stefanie J Krauth, Lewis Steell, Sayem Ahmed, Emma McIntosh, Grace O Dibben, Peter Hanlon, Jim Lewsey, Barbara I Nicholl, David A McAllister, Susan M Smith, et al. 2024. Association of latent class analysis-derived multimorbidity clusters with adverse health outcomes in patients with multiple long-term conditions: comparative results across three UK cohorts. *eClinicalMedicine* 74 (2024).

[6] Valerie Kuan, Spiros Denaxas, Arturo Gonzalez-Izquierdo, Kenan Direk, Osman Bhatti, Shanaz Husain, Shailen Sutaria, Melanie Hingorani, Dorothea Nitsch, Constantinos A Parisinos, et al. 2019. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health* 1, 2 (2019), e63–e77.

[7] Valerie Kuan, Spiros Denaxas, Praveetha Patalay, Dorothea Nitsch, Rohini Mathur, Arturo Gonzalez-Izquierdo, Reecha Sofat, Linda Partridge, Amanda Roberts, Ian CK Wong, et al. 2023. Identifying and visualising multimorbidity and comorbidity patterns in patients in the English National Health Service: a population-based study. *The Lancet Digital Health* 5, 1 (2023), e16–e27.

[8] Claudia Langenberg, Aroon D Hingorani, and Christopher JM Whitty. 2023. Biological and functional multimorbidity—from mechanisms to management. *Nature Medicine* 29, 7 (2023), 1649–1657.

[9] Clare MacRae, Stewart W Mercer, David Henderson, Megan McMinn, Daniel R Morales, Emily Jefferson, Ronan A Lyons, Jane Lyons, Chris Dibben, David A McAllister, et al. 2023. Age, sex, and socioeconomic differences in multimorbidity measured in four ways: UK primary care cross-sectional analysis. *British Journal of General Practice* 73, 729 (2023), e249–e256.

[10] Franck Mauvais-Jarvis, Noel Bairey Merz, Peter J Barnes, Roberta D Brinton, Juan-Jesus Carrero, Dawn L DeMeo, Geert J De Vries, C Neill Epperson, Ramaswamy Govindan, Sabra L Klein, et al. 2020. Sex and gender: modifiers of health, disease, and medicine. *The Lancet* 396, 10250 (2020), 565–582.

[11] Allan L McCutcheon. 1987. *Latent class analysis*. Sage Publications, Beverly Hills, CA.

[12] Sacha Morin, Robin Legault, Zsuzsa Bakk, Charles-Édouard Giguère, Roxane de la Sablonnière, and Éric Lacourse. 2023. StepMix: a python package for pseudo-likelihood estimation of generalized mixture models with external variables. *arXiv preprint arXiv:2304.03853* (2023).

[13] Trishan Panch, Jonathan Pearson-Stuttard, Felix Greaves, and Rifat Atun. 2019. Artificial intelligence: opportunities and risks for public health. *The Lancet Digital Health* 1, 1 (2019), e13–e14.

[14] Alexandra Prados-Torres, Amaia Calderón-Larrañaga, Jorge Hancco-Saavedra, Beatriz Poblador-Plou, and Marjan van den Akker. 2014. Multimorbidity patterns: a systematic review. *Journal of Clinical Epidemiology* 67, 3 (2014), 254–266.

[15] Clark D Russell, Nazir I Lone, and J Kenneth Baillie. 2023. Comorbidities, multimorbidity and COVID-19. *Nature Medicine* 29, 2 (2023), 334–343.

[16] Søren T Skou, Frances S Mair, Martin Fortin, Bruce Guthrie, Bruno P Nunes, J Jaime Miranda, Cynthia M Boyd, Sanghamitra Pati, Sally Mtenga, and Susan M Smith. 2022. Multimorbidity. *Nature Reviews Disease Primers* 8, 1 (2022), 48.

[17] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 17, 3 (2020), 261–272.

[18] Achim Wolf, Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman, and Puja Myles. 2019. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International Journal of Epidemiology* 48, 6 (2019), 1740–1740g.

[19] Sang-Wook Yi, Sangkyu Park, Yong-ho Lee, Hyang-Jeong Park, Beverley Balkau, and Jee-Jeon Yi. 2017. Association between fasting glucose and all-cause mortality according to sex and age: a prospective cohort study. *Scientific Reports* 7, 1 (2017), 8194.