Discovering Prevalent Chronic Disease Profiles with Advanced Clustering Methods

Jiani Yan[1,2,3]

[1]Leverhulme Centre for Demographic Science, University of Oxford
[2]Department of Sociology, University of Oxford
[3]Max Planck Institute for Demographic Research

Abstract

Chronic non-communicable diseases (NCDs) have become pervasive, representing a significant public health challenge in the United Kingdom, particularly due to rising multimorbidity among older populations. Existing research often lacks a holistic, population-level perspective on multimorbidity patterns and associated social determinants. Using the UK Biobank dataset, this study employs a bottom-up analytical approach combining Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering (HDBSCAN) to explore chronic disease patterns across 316 chronic conditions. Nine distinct multimorbidity clusters were identified, characterised by various dominant conditions: Healthy, Asthma, Hypertension, Allergic Rhinitis, Hypertension+Respiratory, Depression, Other Prevalent Diseases, and Heavy Cardiovascular Disease (CVD). Subsequent analysis of social determinants revealed significant differences between healthy and disease-affected groups, notably influenced by age, gender, and drinking behaviours. Clusters such as Heavy CVD showed particularly high disease burdens among older, predominantly male populations, while mental health-related clusters like Depression+ were closely linked with adverse psychosocial factors. This integrated approach highlights the necessity of considering comprehensive social determinants alongside detailed multimorbidity profiles to inform targeted public health interventions and personalised clinical practices.

Keywords: Social Determinants of Health, Clustering, Chronic Diseases

For Correspondence: Jiani Yan, Email: jiani.yan@sociology.ox.ac.uk. **Code Availability**: Codes can be found at GitHub. Please see the readme.md file within that repository for a data availability statement.

Introduction

Chronic non-communicable diseases (NCDs) have emerged as a critical public health concern in the United Kingdom, accounting for a substantial proportion of morbidity and mortality. Like other high-income countries, the UK faces significant health burdens predominantly from cardiovascular diseases, cancers, diabetes, chronic respiratory diseases, and mental health disorders. Recent evidence suggests that multimorbidity—the coexistence of two or more chronic diseases—is increasingly prevalent among the UK population. According to Stafford et al. (2018), over one-quarter of English adults live with at least two chronic health conditions, highlighting a notable rise in multimorbidity. Projections indicate that by 2035, conditions such as arthritis and hypertension will individually affect more than half of adults aged 65 or older in England (approximately 63% and 56%, respectively), with diabetes or cancer each affecting nearly one-quarter of the older population (Kingston et al., 2018). These trends reflect both medical advancements, allowing individuals to live longer with chronic illnesses, and the associated healthcare challenges posed by extended periods of morbidity.

Despite the growing recognition of multimorbidity, current literature predominantly investigates specific subsets of chronic diseases, thus failing to offer a comprehensive view of holistic disease co-occurrence patterns. For example, Launders et al. (2022) utilised primary care data from the UK to examine multimorbidity burdens involving 23 selected diseases, focusing particularly on participants with severe mental illness. Although these 23 diseases were based on established indices such as the Charlson and Elixhauser comorbidity measures with mental illness-specific adjustments, this approach provides only a limited perspective centred around mental health, limiting generalizability to broader multimorbidity patterns. Similar limitations are evident in other studies; for instance, Marengoni et al. (2009) analysed patterns based on a selection of 15 chronic conditions, and Zemedikun et al. (2018) applied clustering methods to the UK Biobank data, considering only 36 pre-selected chronic conditions. Those multimorbidity studies emphasize isolated disease groups or predefined disease categories, thereby neglecting the broader interconnectedness of chronic conditions. This narrow focus prevents the identification of complex disease clusters that might inform more effective healthcare strategies and policy interventions.

Furthermore, the integration of social determinants of health (SDH) in multimorbidity research remains limited. While it is well-established that socioeconomic and environmental factors such as poverty, education, housing, lifestyle behaviors, and social isolation significantly influence disease onset and progression, most studies have explored only a limited scope of these determinants. A recent systematic review by Álvarez-Gálvez et al. (2023) highlighted that research frequently adjusts only for basic demographics like age and sex, with few studies examining an extensive array of social determinants or their interactions with disease clusters. The authors advocate for comprehensive incorporation of social and behavioral determinants into multimorbidity studies, which can elucidate why certain disease clusters disproportionately affect specific populations and identify critical points for intervention. Based on Ng et al. (2018), earlier multimorbidity clustering research typically employed traditional statistical methods such as K-Means clustering, latent class analysis and Multiple Correspondence Analysis (examples see Krauth et al., 2024, Nichols et al., 2022 and Launders et al., 2022). Although widely utilized, these techniques often inadequately capture the complexity inherent in multimorbidity patterns. K-Means clustering, for instance, necessitates specifying the number of clusters beforehand and uses Euclidean distance, which may not adequately represent binary

disease data. LCA, while useful for identifying latent groupings, requires strict probabilistic assumptions that may not accurately reflect the underlying data structure. Consequently, there is growing recognition that traditional methods are insufficiently flexible to represent the complex, nonlinear, and non-additive interactions characteristic of multimorbidity.

Recent advances in data science have led to more sophisticated methodologies capable of uncovering complex multimorbidity structures. Graph-based community detection algorithms and manifold learning methods are increasingly being employed to model disease clusters without the restrictive assumptions of traditional methods. For instance, Beaney et al. (2024) successfully employed the Markov Multiscale Community Detection (MMCD) method in a comprehensive network analysis of primary care data covering 253 diseases in England, demonstrating its potential for uncovering nuanced disease clusters. Despite the methodological sophistication and the detailed comparisons among various clustering algorithms, their study falls short in providing sufficient interpretability and clinical context. Specifically, the outcomes were inadequately adjusted or explained using basic descriptive statistics such as disease prevalence within each identified cluster. Moreover, the absence of discussions surrounding fundamental social determinants of health, including age and gender, further limits the interpretability and comparability of their findings with other multimorbidity research. This omission emphasizes the need for a holistic analytical approach that integrates social perspectives alongside advanced methodological frameworks to enhance relevance and applicability in multimorbidity studies.

Similarly, Jiang et al. (2025) employed advanced machine learning algorithms, including XGBoost and Neural Networks, integrating multimodal information from health records, genetic data, and medical imaging to predict disease outcomes. Particularly, their neural network approach achieved the highest predictive performance by jointly modeling multiple diseases, allowing them to construct a disease distance matrix from the weight matrices preceding the output layer. This enabled a comprehensive mapping of multimorbidity patterns using an extensive list of 1560 diseases. Despite the innovative methodological approach, the derived disease relationships depend heavily on the predictive accuracy and generalizability of the neural network model. Therefore, the representativeness of the constructed distance matrix remains contingent upon model performance, potentially limiting its authenticity in capturing true disease interconnectedness. Such technical complexity alone does not guarantee accurate representation of disease co-occurrence if the model fails to adequately reflect real-world disease associations inherent in the original data.

Building on these methodological advancements, our study employs a stepwise analytical approach using UMAP and HDBSCAN to identify disease clusters within the UK Biobank dataset. UMAP effectively reduces high-dimensional multimorbidity data into a low-dimensional representation while preserving both local and global structural relationships. Subsequently, HDBSCAN identifies clusters of arbitrary shapes and explicitly manages noise points, addressing limitations inherent to traditional methods such as K-Means and hierarchical clustering. This sophisticated approach captures complex, nonlinear interrelationships among chronic diseases, thereby revealing nuanced clusters defined by intricate condition interactions that might be overlooked by conventional techniques. Additionally, HDBSCAN autonomously determines the optimal number of clusters, significantly enhancing the robustness and interpretability of clustering results.

In addition to disease clustering, we utilise comprehensive survey data available in the

UK Biobank dataset—including demographic, socioeconomic, behavioural, and mental health information—to investigate differences across identified clusters from a social determinants perspective. This post-clustering analysis enables us to gain deeper insights into the social context and underlying factors associated with each multimorbidity cluster (refer to Figure 1 for detailed research workflow). Overall, this study aims to deliver an integrative and holistic understanding of multimorbidity patterns, emphasising the critical role of extensive social determinants data to inform targeted public health interventions and enhance clinical practice.

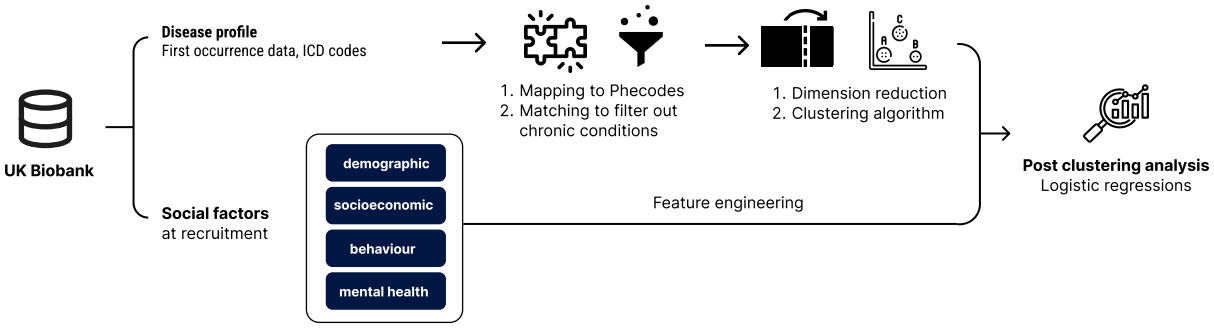

Figure 1: **Schematic Workflow Plot of Chapter Two**

Data

To preserve the granularity and breadth of information on both diseases and social determinants, this study utilizes data from the UK Biobank (Sudlow et al., 2015), a large-scale prospective cohort comprising approximately 500,000 participants (502,129 in our study) aged 40 to 69 years at recruitment. The UK Biobank offers a rich repository of de-identified genetic, lifestyle, and longitudinal health records, along with biological samples, making it an unparalleled resource for population health research. A notable feature of the UK Biobank is its integration of multi-source clinical records through the *first occurrence* data schema, which captures the earliest documented diagnosis of health conditions per participant. These events are aggregated from diverse sources, including primary care records, hospital inpatient admissions, and self-reported medical histories. To further enhance diagnostic completeness, we manually integrated cancer diagnoses from the linked National Cancer Registry, enabling a more exhaustive representation of disease profiles.

In this study, we draw on social factors measured at recruitment, as this wave provides the most comprehensive coverage of both social determinants and participants. These factors are grouped into four domains: demographic, socioeconomic, health behaviours, and mental health. Each domain encompasses distinct variables that together offer a multifaceted perspective on participants' social environments. Incorporating multiple domains allows us to capture a more holistic view of how disease clusters emerge and to better understand the characteristics underlying their formation. Specifically, the demographic domain includes age and gender; the socioeconomic domain comprises education qualifications, job code, total household income, and home ownership; the health behaviours domain covers seven indicators, including current and past alcohol consumption, current and past smoking, insomnia, and vigorous and moderate physical activity; and the mental health domain incorporates unpleasant feelings (covering fed

up, guilty, nervous, sensitive, and anxious feelings) and loneliness. A statistical description of these social factors is provided in Table 1.

Domain	Social factor	Mean	Standard deviation
Demographic	Male	0.46	0.50
	Age at recruitment	56.53	8.09
Socioeconomic	Education qualifications	3.89	1.28
	Job code	6.62	1.84
	Total household income	2.66	1.11
	Renting home	0.92	0.67
	Townsend deprivation index	-1.30	3.09
Health behaviours	Current Alcohol Drinker	0.92	0.28
	Ever Drink Alcohol	0.96	0.21
	Current tobacco smoking	0.18	0.55
	Ever smoked	0.60	0.49
	Insomnia	2.04	0.72
	Vigorous physical activity	1.79	1.91
	Moderate physical activity	3.59	2.27
Mental health	Unpleasant Feelings	2.04	1.56
	Loneliness, isolation	0.19	0.38

Table 1: Statistical description of included social factors

The UK biobank mainly provides disease information in the ICD-10 format, however, since ICD-10 codes are known for being often fragmented and overly granular for epidemiological analysis, we map diagnostic information to Phecodes using the established mapping framework by Wu et al. (2019). Phecodes group related ICD codes into clinically meaningful phenotypes, reducing dimensional complexity while preserving diagnostic relevance. For example, the Phecode 495.0 for Asthma maps five closely related asthma ICD-10 codes together(including J45 Asthma, J45.0 Predominantly allergic asthma, J45.1 Nonallergic asthma, J45.8 Mixed asthma, and J45.9 Asthma, unspecified). This harmonisation is especially pertinent in our study, where fragmentation in ICD coding can obscure disease patterns with too refined granularity. Please see Table S1 for the mapping between the Phecodes and ICD-10 codes for the top ten prealent chronic diseases in the UKBB. To focus on enduring health burdens, we restrict our disease set to those classified as chronic conditions. Chronic conditions are defined as diseases that are typically persistent, require long-term management, and are rarely resolved spontaneously. We adopt the classification from the Chronic Condition Indicator Refined (CCIR) tool (Agency for Healthcare Research and Quality, 2024), which systematically categorizes ICD-10-CM codes into chronic and non-chronic conditions based on clinical severity, duration, and need for continuous treatment.

Since the CCIR tool operates on ICD-10-CM codes — whereas the UK Biobank primarily uses ICD-10 — we implemented a cross-mapping strategy. This leveraged the dual ICD-10 and ICD-10-CM mappings available in the Phecode system to serve as an intermediary. Specifically, each ICD-10 code in the UK Biobank data was first mapped to a Phecode; we then used the corresponding ICD-10-CM mappings associated with each Phecode to identify whether it falls under the chronic condition category. This two-step approach enabled consistent and

clinically informed classification of chronic conditions across disparate coding systems.

Method

Here we discuss the methods to perform dimension reduction and clustering: the UMAP and HDBSCAN. In high-dimensional spaces, distance metrics become less meaningful, and density-based clustering algorithms like HDBSCAN struggle to identify clusters effectively. By reducing the data to a lower-dimensional space using UMAP, we preserve the essential topological features necessary for effective clustering. This preprocessing step enhances the performance and accuracy of HDBSCAN, as it operates more effectively in lower-dimensional spaces where density estimates are more reliable (Dalmia & Sia, 2021).

UMAP

Given the high dimensionality of our disease data and to mitigate the 'curse of dimensionality' inherent in high-dimensional datasets, we plan to apply the UMAP algorithm prior to the clustering step. UMAP is a nonlinear dimensionality reduction technique rooted in manifold learning and topological data analysis. It is designed to preserve both local and global data structures during the embedding process, via constructing a graphical representation of the data topology and optimising a low-dimensional embedding to preserve the topological structure (McInnes et al., 2020).

UMAP is based on the assumption that the data lies on a Riemannian manifold $\mathscr{M}$ and locally approximates the manifold via a fuzzy topological structure. Given a dataset $\mathscr{X} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^D$ with n observations and D input features, it constructs a weighted k-nearest neighbor (k-NN) graph based on the Euclidean distance in the original space. The edge weights between two data points i, j are computed as:

$$\mu_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \tag{1}$$

where $d(x_i, x_j)$ is the distance between points x_i and x_j. ρ_i is the distance to the nearest neighbour of x_i to ensure its local connectivity. σ_i is a local connectivity parameter, chosen such that the Shannon entropy of the connectivity distribution equals a fixed target:

$$\sum_{j=1}^{k} \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) = \log_2(k) \tag{2}$$

where k is the number of neighbours determined by the hyperparameter `n_neighbors`. Then the fuzzy topological structure in the high-dimensional space is symmetrised using:

$$P_{ij} = \mu_{ij} + \mu_{ji} - \mu_{ij} \cdot \mu_{ji} \tag{3}$$

UMAP seeks a low-dimensional embedding $\mathscr{Y} = \{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^d$, where the dimension of the low-dimension embedding d is set by the `n_components` hyperparameter. In the low-dimensional space, a similar fuzzy set Q_{ij} is defined using a smooth approximation to the 1-simplex:

$$Q_{ij} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1} \tag{4}$$

where y_i, y_j are the embedded low-dimensional points and a, b are parameters controlling the tightness of the embedding, determined from the `min_dist` hyperparameter. It controls the minimum allowed distance between points in the low-dimensional embedding, where a higher value results in more compact clusters with fine-grained local structure preserved. The embedding is optimised by minimizing the cross-entropy between the high-dimensional and low-dimensional fuzzy simplicial sets P and Q:

$$C = \sum_{(i,j)} P_{ij} \log \frac{P_{ij}}{Q_{ij}} + (1 - P_{ij}) \log \frac{1 - P_{ij}}{1 - Q_{ij}} \tag{5}$$

UMAP is particularly suitable for our dataset due to its computational scalability ($\mathscr{O}(N \log N)$) and its ability to preserve both local and global structure, which is critical when working with large, high-dimensional data matrices such as ours. In this paper, only the chronic diseases at the phecode level of all observations will be sent to the UMAP algorithm, since the research aim of this work is to capture the patterns at the disease level.

HDBSCAN

HDBSCAN is an unsupervised clustering algorithm that extends Density-Based Spatial Clustering of Applications with Noise (DBSCAN) by converting it into a hierarchical clustering algorithm and then using a technique to extract a flat clustering based on the stability of clusters. It is particularly adept at identifying clusters of varying densities and arbitrary shapes, and it effectively handles noise in the data (McInnes & Healy, 2017).

Since HDBSCAN builds on the DBSCAN method, we will introduce DBSCAN first and then HDBSCAN. Introduced by Ester et al. (1996), DBSCAN is a foundational density-based clustering algorithm that identifies clusters as contiguous regions of high point density, separated by regions of low density. Given a dataset $\mathscr{X} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d$ (in our case, they are the lower-dimension embeddings generated from UMAP in the last step), DBSCAN categories any point x_i as core point, border point or noise point based on the following function of the $\varepsilon-$neighborhood of x_i:

$$N_\varepsilon(x_i) = \{x_j \in \mathscr{X} \mid \|x_i - x_j\| \leq \varepsilon\} \tag{6}$$

where ε is the hyperparameter of the minimum radius defining the neighborhood around a point. x_i together with all the points in the neighborhood $N_\varepsilon(x_i)$ are core points if $|N_\varepsilon(x_i)| \geq$

`min_cluster_size`, which is the hyperparameter defining the minimum number of points for a group to be considered a cluster. Border points are not core points themselves, but within the ε-neighborhood of a core point. Noise points are the points neither core nor border points. Clusters are then formed by connecting core points and their reachable border points into maximally connected components under density-reachability. However, DBSCAN's reliance on a global ε makes it unsuitable for datasets with variable density. Additionally, it does not provide a multiscale view of the data, limiting its adaptability in exploratory analyses.

These challenges of DBSCAN motivated the development of HDBSCAN, which extends DBSCAN by removing the requirement for a global density threshold and instead builds a hierarchy of clusters based on varying density levels. It retains DBSCAN's core advantages such as noise detection and arbitrary cluster shapes while improving robustness and flexibility. The algorithm introduces the core distance for each point x_i, defined using the user-defined hyperparameter `min_samples`: $\text{core}_m(x_i) = d(x_i, x(m))$ where $x_{(m)}$ is the m-th nearest neighbor of x_i. This forms the basis for the mutual reachability distance between any pair of points x_i and x_j:

$$d_{\text{mreach}}(x_i, x_j) = \max\left(\text{core}_m(x_i), \text{core}_m(x_j), d(x_i, x_j)\right) \tag{7}$$

This transformation normalises local density variations and ensures that all clusters are comparable on a common density scale. HDBSCAN then builds a complete weighted graph, from which it extracts a Minimum Spanning Tree (MST) that encodes the essential structure of the data's density connectivity by building a hierarchiccal cluster tree. The MST removes edges in ascending order of the mutual reachability distance. As the distance threshold increases, clusters split or dissolve. The resulting tree encodes a hierarchy of clusterings at multiple density levels. The hierarchy is then built by removing edges from the MST in order of decreasing mutual reachability distance, creating a dendrogram. A flat clustering is extracted from the hierarchy by selecting the most stable clusters, defined by the following stability score:

$$\text{Stability}(C) = \int_{\lambda_{\min}}^{\lambda_{\max}} |C_\lambda| \, d\lambda \tag{8}$$

where $\lambda = \frac{1}{d_{\text{mreach}}}$ is the inverse of mutual reachability, and $|C_\lambda|$ is the size of the cluster at scale λ.

Hyper-parameter Optimisation

. The hyperparameters for UMAP and HDBSCAN are optimised through a grid search strategy that evaluates combinations of parameters based on a composite objective function. This objective function maximises the harmony between three quality metrics computed from the clustering results: silhouette score, persistence score and noise rate.

The silhouette score quantifies the cohesion and separation of the resulting clusters. For a given data point i, it is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

where $a(i)$ is the mean intra-cluster distance (i.e., the average distance to all other points in the same cluster), and $b(i)$ is the mean nearest-cluster distance (i.e., the average distance to all points in the nearest cluster not containing i). The overall silhouette score is the mean of $s(i)$ across all clustered points.

Persistence, a metric derived from HDBSCAN's hierarchical clustering tree, reflects the robustness of each cluster. The persistence of a cluster C is given by:

$$\text{Persistence}(C) = \lambda_{\text{end}}(C) - \lambda_{\text{start}}(C) \tag{10}$$

where $\lambda_{\text{start}}(C)$ and $\lambda_{\text{end}}(C)$ represent the inverse mutual reachability distances at which cluster C first appears and later disappears in the cluster hierarchy. The mean persistence across all selected clusters serves as a measure of stability.

Noise rate measures the proportion of points labelled as noise (i.e., not assigned to any cluster) by HDBSCAN. Let N be the number of data points and N_{noise} the number of noise-labelled points. Then,

$$\text{Noise Rate} = \frac{N_{\text{noise}}}{N} \tag{11}$$

A lower noise rate generally indicates better clustering coverage, although some noise is expected and can be desirable when separating outliers.

The final selection criterion is a weighted composite function:

$$\mathscr{L} = \alpha \cdot \bar{s} + \beta \cdot \bar{p} - \gamma \cdot r \tag{12}$$

where $\bar{s}$ is the mean silhouette score, $\bar{p}$ is the mean persistence, r is the noise rate, and $\alpha, \beta, \gamma \in \mathbf{R}^+$ are weights calibrated to balance the three metrics. The optimal hyperparameters are selected by maximising $\mathscr{L}$ over the grid. In this paper, we set the α, β and γ to 0.5, 0.25 and 0.25, respectively, to reflect a higher emphasis on the silhouette score. The grid searching space consists of seven main hyper-parameters: the `n_component`, `n_neighbors`, `random_state`, `min_dist`, `min_cluster_size`,`min_cluster` and `min_samples`.

Results

In this section, we present the findings from our clustering analysis, structured into two primary parts. Initially, we detail the clustering outcomes derived from our UMAP+HDBSCAN analytical pipeline. This unsupervised approach facilitates the detection of distinct patterns of chronic

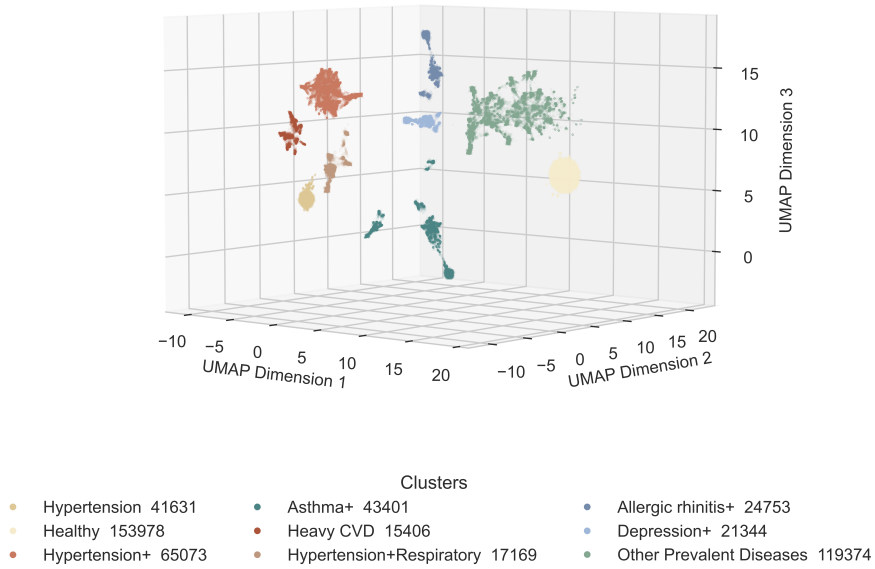

Figure 2: Distribution of disease clusters identified by UMAP+HDBSCAN

disease co-occurrence, offering a comprehensive overview of disease burden and multimorbidity profiles within the study population. Subsequently, we investigate associations between social determinants and the identified disease clusters. We compare the sociodemographic, behavioural, and psychological characteristics of individuals across clusters, emphasising contrasts between each disease cluster and the healthy baseline group. Our objective is to identify social factors linked to specific clusters, highlighting both the characteristics that distinguish individuals experiencing poorer health from those in good health, as well as the diversity in social profiles across different disease-afflicted groups.

Clustering Results

Here, we report the results of the UMAP+HDBSCAN clustering analysis. A robustness check of the concordance across different choices of weights (α, β, and γ) is shown in Figure S2. The selected set of hyperparameters yielded a clustering quality of $\mathscr{L} = 0.427$. We chose this configuration because it achieves comparable high clustering quality while providing a finer-grained partition of the data, enabling us to explore differences not only in disease composition but also in the social heterogeneity between clusters (see Figure S3 for the distribution of $\mathscr{L}$ scores from the grid search and a comparison of clustering results). Using these hyperparameters, we identified nine distinct clusters, illustrated in Figure 2. Healthy, Hypertension, Hypertension+, Hypertension+Respiratory, Heavy CVD, Asthma+, Allergic Rhinitis+, Depression+, and Other Prevalent Diseases. Among these clusters, the largest groups were Other Prevalent Diseases (119,374 individuals), Healthy (153,978), and Hypertension+ (65,073). Smaller clusters (fewer than 50,000 observations), such as Hypertension+, Hypertension+Respiratory, and Heavy CVD, could potentially merge into a larger group under less stringent HDBSCAN

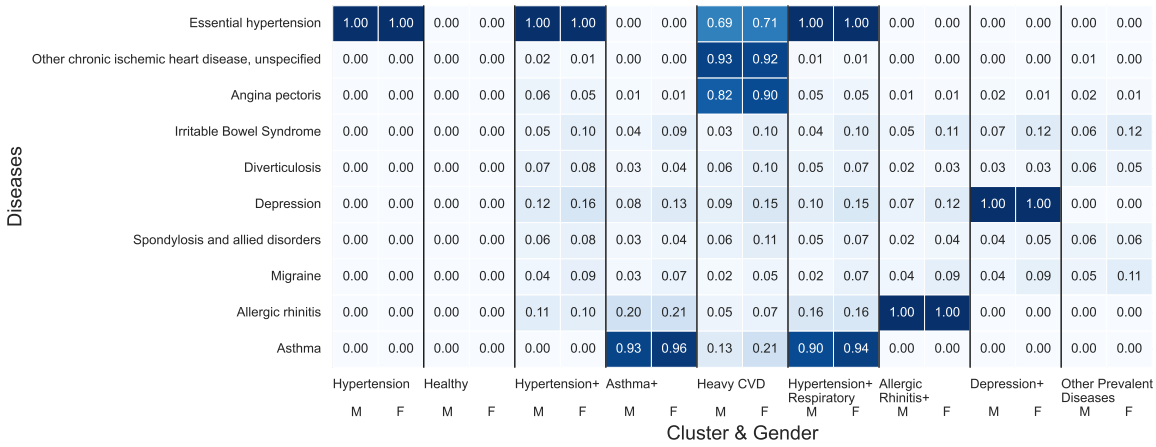

Figure 3: Prevalence of top 10 Phecode diseases in each cluster by gender

clustering conditions. Similarly, Allergic Rhinitis+, Depression+, and Asthma+ could also aggregate into a larger cluster. However, to facilitate a detailed exploration of the associations of social factors and provide more granular interpretations, we retained these finer distinctions. Figure 3 displays the prevalence of the top ten chronic diseases across the nine clusters, stratified by gender. Higher prevalence values indicate a greater incidence of specific diseases within each cluster and gender group. For instance, both men and women uniformly exhibit allergic rhinitis within the Allergic Rhinitis+ cluster, as is the case for depression in the Depression+ cluster. From this visualisation, it is evident that most clusters have clearly dominant conditions, except for the Healthy and Other Prevalent Diseases clusters.

Table 2: Cluster Summary Statistics

Cluster	Mean No. Disease	Age	Male	Min No. Disease	Household Income
Healthy	0.03	54.46	0.48	0	2.85
Hypertension	1.09	58.68	0.56	1	2.60
Other Prevalent Diseases	1.65	56.99	0.38	1	2.64
Allergic rhinitis+	2.02	53.92	0.44	1	2.88
Depression+	2.19	54.62	0.31	1	2.44
Asthma+	2.39	54.72	0.41	1	2.73
Hypertension+	3.13	60.10	0.50	2	2.39
Hypertension+Respiratory	3.96	58.88	0.46	2	2.41
Heavy CVD	5.27	62.21	0.77	2	2.17

Table 2 summarises key statistics of the nine clusters identified in our analysis, including mean disease count, mean age, male proportion, and minimum disease count. The Hypertension and Other Prevalent Diseases clusters both exhibit an average disease count below 2, with the Hypertension cluster comprising 56% males and the Other Prevalent Diseases cluster consisting of only 38% males. The Heavy CVD cluster shows the highest average disease count (5.27) and represents the oldest demographic group among the clusters. It is notably male-dominated, with males accounting for 77% of the group. Regarding minimum disease counts, the Healthy cluster exclusively includes participants without chronic diseases. Clusters such as Allergic Rhinitis+, Depression+, and Asthma+ indicate multimorbidity, reflected by their average disease counts exceeding two. Nevertheless, these clusters still include some individuals with only the specific conditions indicated by their cluster names. Clusters labelled Hyperten-

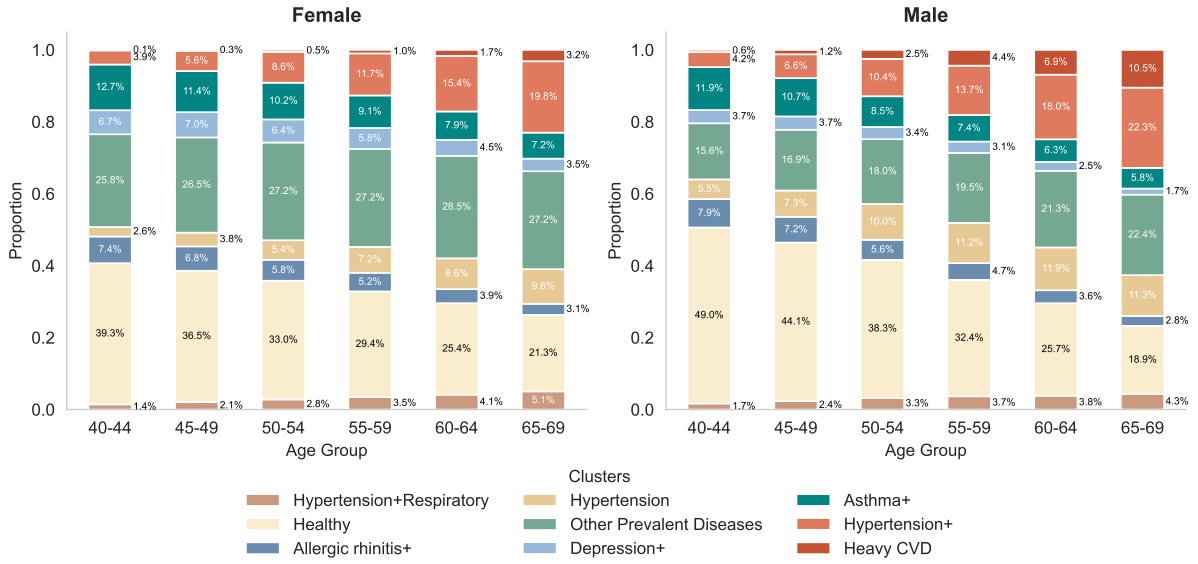

Figure 4: Age- and gender-specific distribution of multimorbidity clusters.

sion+, Hypertension+Respiratory, and Heavy CVD distinctly represent multimorbid groups, where each participant has at least two chronic diseases. Notably, the Depression+ cluster has the highest female representation, comprising 69% females.

Figure 4 displays the proportional distribution of clusters across age groups for women and men. Several ageing-related trends are consistent across both genders. Most notably, the proportion classified as Healthy declines steadily with advancing age. Among women, the Healthy cluster decreases from 39.3% at ages 40–44 to 21.3% at ages 65–69, while among men the corresponding decline is even steeper, from 49.0% to 18.9%. This pattern illustrates the expected erosion of 'healthy' status with age and the progressive accumulation of multimorbidity across the life course. At the same time, disease-related clusters rise in prevalence with age. In women, the Hypertension+ cluster grows from 3.3% at 40–44 to nearly one in five (19.8%) by 60–64, and 19.8% remains prevalent in 65–69. Similarly, in men the Hypertension+ cluster increases from 6.6% at 40–44 to 22.3% at 65–69. The Heavy CVD cluster also becomes more prominent with age, particularly in men, where it reaches 4.3% by 65–69 compared with 5.1% in women, reflecting the well-known higher cardiovascular burden in males. Taken together, these findings underscore that while both sexes accumulate cardiovascular multimorbidity with age, the magnitude of the increase is greater in men.

Gender-specific differences are particularly evident when examining mental health–related multimorbidity. Across all age groups, women exhibit a higher proportion of the Depression+ cluster compared to men. For instance, in the 40–44 age group, Depression+ is 7.4% in women versus 7.9% in men, but as age advances, the gender gap becomes more pronounced: by 65–69, Depression+ affects 9.6% of women compared with only 5.8% of men. This suggests that depression-associated multimorbidity is disproportionately borne by women in older adulthood, consistent with prior epidemiological evidence on gendered patterns of mental health. Despite these differences, similarities also exist. The Other Prevalent Diseases cluster maintains a stable and substantial share across sexes and age groups, accounting for roughly one quarter of individuals in later adulthood (27.2% in women and 22.4% in men aged 65–69). Likewise, clusters such as Allergic rhinitis+ and Asthma+ remain relatively minor contribu-

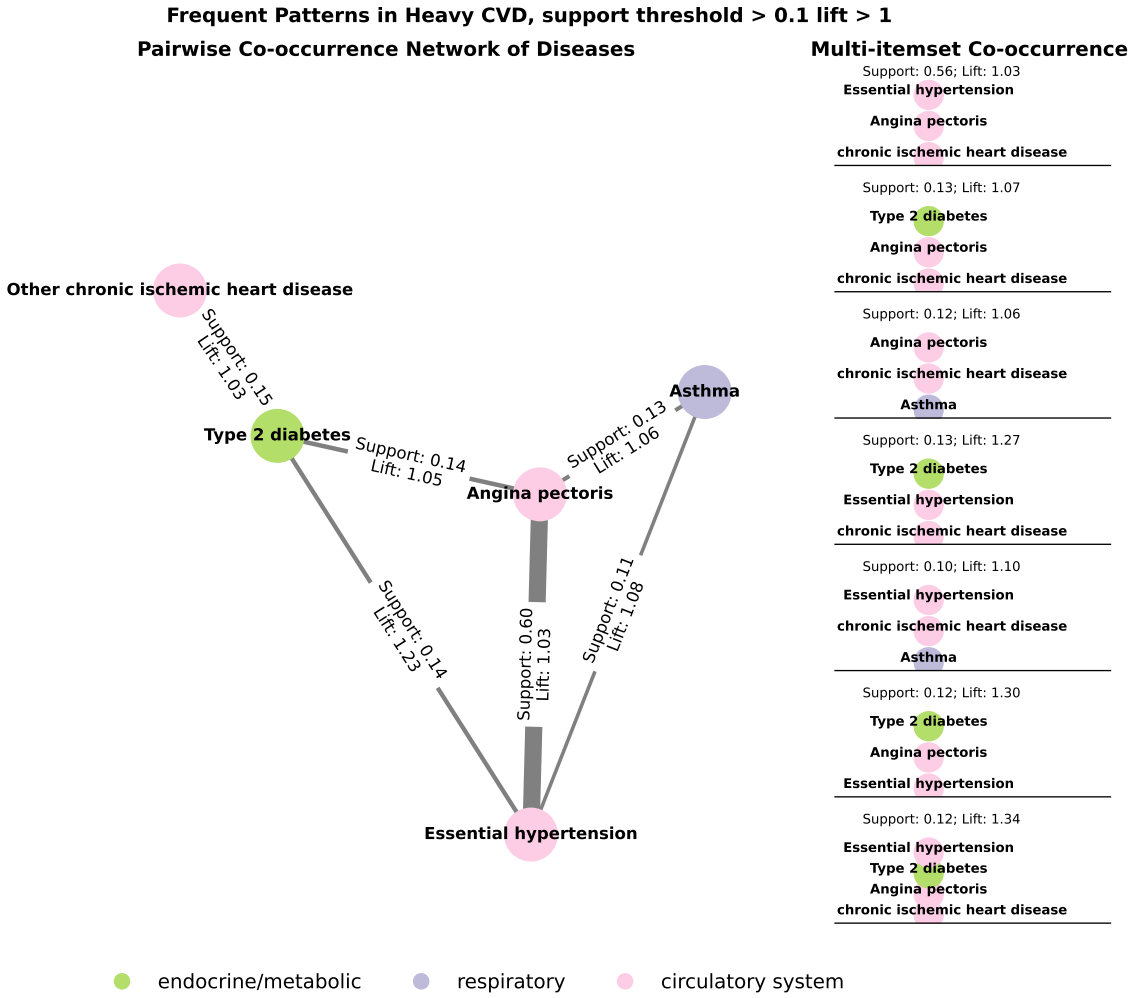

Figure 5: Network View of Frequent Co-occurrence Patterns in Heavy CVD Cluster, where support threshold > 0.1 and lift > 1

tors across all strata, though women show slightly higher proportions of Asthma+ in early and mid-life (12.7% at 40–44 vs. 11.9% in men), whereas men maintain slightly higher shares of Allergic rhinitis+ (7.9% at 40–44 vs. 7.4% in women). These smaller clusters are stable in prevalence and do not appear to drive major gender differences.

We also identify frequent disease co-occurrence patterns within each disease cluster by applying support and lift thresholds tailored to each cluster's characteristics. Here, support is defined as the proportion of the dataset in which a specific itemset appears, reflecting how frequently certain combinations of diseases occur together. Lift measures the strength of association between diseases, calculated as the ratio of the observed support to the expected support assuming diseases are independent. An itemset represents a combination or group of diseases that frequently appear together within the dataset. For example, the Heavy CVD cluster exhibits the highest average disease count, with a mean of 5.27 diseases per individual. Figure 5 displays prevalent disease co-occurrence patterns selected based on a minimum support threshold of 10% among all frequent itemsets. The first row in the figure illustrates networks of disease co-occurrences involving three or more conditions, while the second row presents pairwise co-occurrence relationships. All patterns depicted have lift values exceeding 1, signifying that these disease combinations occur more frequently than would be expected by chance alone.

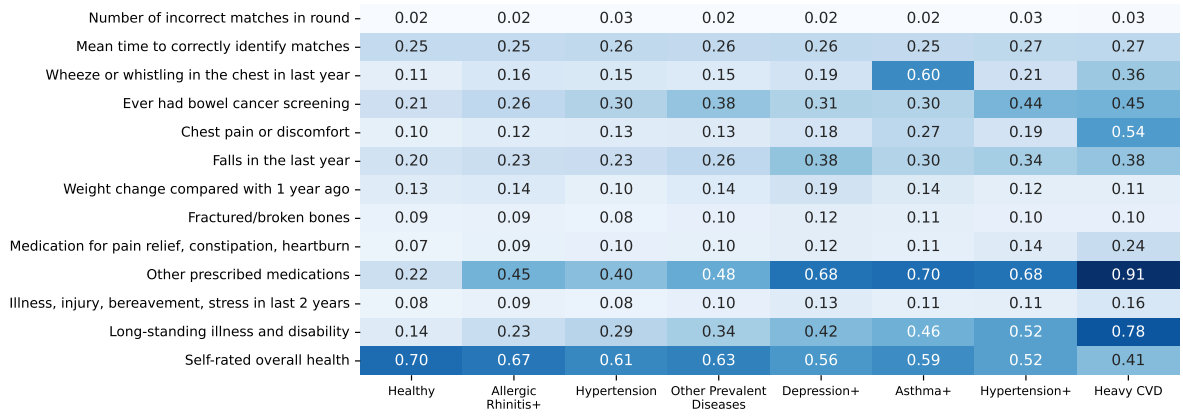

Figure 6: Physical severity by cluster

Within the Heavy CVD cluster, the most prevalent multi-disease combination comprises angina pectoris, essential hypertension, and other chronic ischemic heart disease, with a support level of 56%. The highest pairwise co-occurrence involves essential hypertension and angina pectoris, observed in 60% of cases. Type 2 diabetes also demonstrates notably high co-occurrence rates with several cardiovascular conditions, including other chronic ischemic heart disease (15%), essential hypertension (14%), and angina pectoris (14%). Additional detailed network visualisations for other disease clusters are available in the Supplementary Information section of this Chapter.

Lastly, we discuss the physical severity associated with each cluster, illustrated in Figure 6. Physical health indicators considered include health screening outcomes (e.g., wheezing or chest whistling in the past year, history of bowel cancer screening), symptoms and medication use (e.g., chest pain, recent falls, fractures in the last five years, medication for pain relief, constipation, heartburn, prescription medication usage, presence of long-standing illness or disability), recent changes in health status (e.g., weight changes, recent illness, injury, stress, or bereavement), and self-rated overall health. All variables are scaled between zero and one, where a higher value generally indicates worse physical conditions, except for self-rated overall health, where a higher value signifies better perceived health. Each numerical value represents the cluster mean for the respective variable. In terms of physical health severity, the healthy cluster, as anticipated, exhibits the lowest levels of physical health burden. Detailed comparisons of participants within the healthy group who nonetheless have chronic conditions can be found in Figure S1. Clusters involving hypertension and multiple conditions demonstrate significant physical health challenges, including lower self-rated overall health, higher usage of prescribed medications, greater incidence of pain, and more frequent falls.

The Heavy CVD cluster stands out as particularly severe, with 91% of participants taking prescription medications and 78% reporting long-standing illnesses, disabilities, or infirmities. Asthma-related clusters display notably higher proportions of respiratory symptoms such as wheezing or chest whistling in the past year, with the Asthma+ cluster at 60% and Hypertension+Respiratory cluster at 65%. These rates considerably exceed those observed in other clusters, including Heavy CVD (36%). These findings highlight the unique disease burdens borne by different clusters, closely associated with their designated ‘index’ conditions identified in the cluster names. Although the Healthy cluster generally demonstrates a very low average disease count, a small proportion of its participants still have diseases, primarily rare

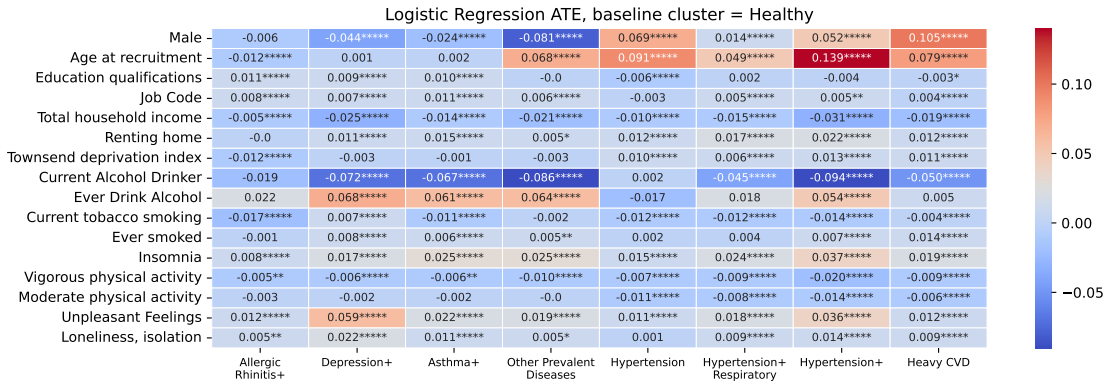

Figure 7: Logistic regression ATE with healthy cluster as the baseline cluster.

cancers that rarely co-occur with other chronic conditions. For clarity in subsequent analyses, especially when comparing social determinants across clusters, we will explicitly distinguish between genuinely healthy participants and those with rare diseases. Further details on the categorisation of rare diseases within the Healthy cluster will be elaborated upon in the supplementary information.

Clusters and Social Factors: Comparing Disease and Healthy Groups

In this section, we examine social factors differentiating disease clusters from the healthy baseline cluster. We apply logistic regression analyses where the outcome variable equals one for participants in a specific disease cluster, zero for those in the healthy cluster, and is undefined (None) for other groups. The healthy cluster, as previously described, comprises participants without chronic diseases. Average Treatment Effects (ATE) for each social factor across all disease clusters are displayed in Figure 7. Binary variables, such as gender (Male), current and ever alcohol drinking status and tobacco smoking habits (current and ever smoked) remain unstandardised, while all other variables are standardised. Hence, the ATE represents the effect of a one-standard-deviation increase in these continuous variables on the likelihood of belonging to each disease cluster versus the healthy group.

Figure 7 uses colour intensity to illustrate the magnitude of effects, with darker shades indicating stronger differentiation from the healthy group. For example, a one-standard-deviation increase in age is associated with a 13.9% higher likelihood of being in the Hypertension+ cluster compared to the healthy cluster. Statistical significance levels are also indicated in the figure: '**' denotes significance at the p-value <0.05 level, while a more stringent criterion of '*****' indicates significance at p <0.00001. All p-values reported have been adjusted for multiple comparisons using the Bonferroni correction due to the large dataset size and multiple variables fitted. Overall, the most substantial differentiators between disease clusters and the healthy cluster are age, gender, and current alcohol consumption. Regarding socioeconomic factors, average total household income before tax is consistently significant across all disease clusters.

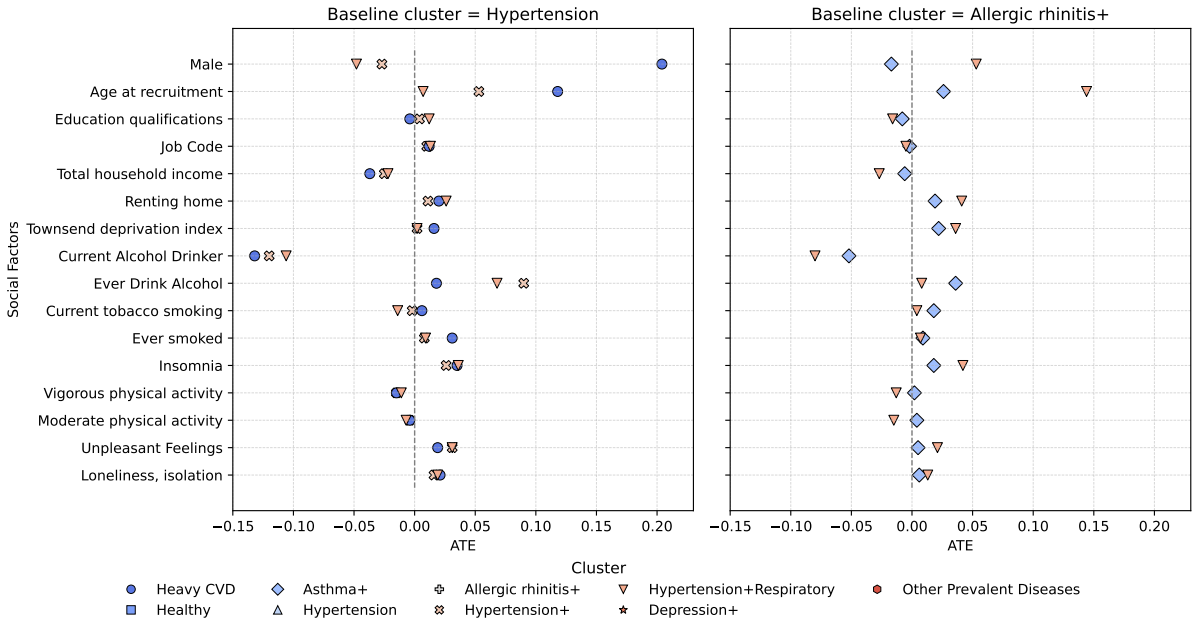

Figure 8: ATE of social factors between hypertension-related clusters and respiratory-related clusters

Clusters with Social factors: Between-Disease Clusters

In this section, we compare the social exposures between the disease clusters, especially between the different hypertension-related clusters and between the other prevalence clusters and the rest disease clusters. The within-hypertension clusters comparsion can disclose the social difference between different type of hypertension related disease clusters. Similarily, we are interested in the social difference between the participants who have respiratory-related diseases.

Four clusters identified in our analysis prominently feature hypertension: Hypertension, Hypertension+, Hypertension+Respiratory, and Heavy CVD. Almost all participants in these clusters exhibit hypertension, except the Heavy CVD cluster, which maintains a prevalence of approximately 70%. Severity indicators such as self-rated overall health and the presence of long-standing illness, disability, or infirmity reveal a clear hierarchy of severity, increasing from Hypertension, Hypertension+, Hypertension+Respiratory, to Heavy CVD. Utilising the Hypertension-only cluster as a reference baseline, Figure 8 illustrates the ATE of various social factors comparing other hypertension-related clusters to this baseline group. Significant social differences emerge between individuals in the hypertension-only cluster and those in the other hypertension-related clusters, notably in terms of gender, age, alcohol consumption, sleep quality, and mental health. Generally, Hypertension+ and Hypertension+Respiratory clusters include a higher proportion of females compared to the baseline hypertension cluster, while the Heavy CVD cluster is more male-dominated. Regarding age, individuals in the Hypertension-only cluster are the youngest, with increasing age across Hypertension+, Hypertension+Respiratory, and Heavy CVD clusters.

Interestingly, all clusters beyond hypertension-only exhibit lower proportions of current alcohol drinkers, an observation that initially appears counterintuitive given the increasing severity. However, since the UK Biobank captures historical disease diagnoses, these patterns might

reflect behavioural changes following disease onset. Supporting this interpretation, historical alcohol consumption ('ever drank alcohol') is higher across these clusters compared to the baseline hypertension-only group, inversely correlating with current drinking patterns. In terms of physical activity, particularly vigorous physical activity, all clusters show an increase in inactivity compared to the hypertension-only group. This could partially result from higher physical constraints in these clusters, as demonstrated by markedly elevated proportions of long-standing illness or disability—for example, 78% in the Heavy CVD cluster versus only 29% in the hypertension-only cluster.

Lastly, we use the Allergic rhinitis+ cluster as the baseline to examine differences in social factors across multimorbidity clusters, illustrated in the right panel of Figure 8. In general, no single social factor uniformly distinguishes the Allergic rhinitis+ cluster from all others. However, several systematic patterns emerge. With respect to gender, clusters with heavier cardiometabolic burdens (e.g., Hypertension+ and Heavy CVD) are more male-dominated compared to the Allergic rhinitis+ cluster, whereas the Depression+ cluster shows an opposite pattern with a higher representation of women. In terms of age, the Allergic rhinitis+ cluster is consistently younger than the hypertension- and CVD-related clusters, but older than clusters such as Asthma+ and Healthy. This intermediate position is further reflected in lifestyle behaviours: for example, Allergic rhinitis+ shows lower probabilities of alcohol consumption compared to Hypertension+ and Healthy, but higher probabilities compared to Depression+ and Asthma+. Smoking behaviour reveals a similar gradient, with Allergic rhinitis+ lying between clusters with heavy tobacco exposure (e.g., Hypertension+Respiratory) and those with lower exposure. Mental health–related social factors differentiate the clusters most strongly. Individuals in the Depression+ cluster exhibit markedly higher ATEs for unpleasant feelings, loneliness, and social isolation compared to the Allergic rhinitis+ cluster, highlighting the strong psychosocial dimension of this disease pattern. Overall, the comparisons suggest that Allergic rhinitis+ occupies a socio-behavioural middle ground: younger and with somewhat healthier lifestyle profiles than cardiometabolic clusters, but distinct from mental health–related multimorbidity patterns, which remain more socially disadvantaged.

Conclusion

In this study, we applied UMAP and HDBSCAN to chronic disease profiles constructed from the UK Biobank first-occurrence records and linked cancer registry data to analyse co-occurrence patterns of diseases agnostically. A key novelty of this work lies in the scale and complexity of the input data: unlike previous studies that pre-selected a limited set of conditions, our analysis incorporated 316 distinct chronic diseases, providing one of the most comprehensive multimorbidity clustering investigations to date. This broad inclusion allows for the discovery of patterns that may otherwise remain hidden when restricting attention to a predefined disease list. The methodological framework constitutes another major innovation. By leveraging UMAP and HDBSCAN, we demonstrate the capacity of these advanced unsupervised learning techniques to capture non-linear and complex relationships among chronic diseases(e.g. shared etiologies and causal dependencies where one condition predisposes to or exacerbates another). This computational capability enabled us to reveal meaningful multimorbidity profiles that extend beyond what can be uncovered using traditional linear clustering methods. Our clustering analysis revealed distinct multimorbidity profiles characterised by varied disease bur-

dens, demographic distributions, and physical severities. The identified clusters illustrate clear contrasts not only in disease composition but also in demographic characteristics. For example, the Heavy CVD cluster exhibited the highest disease burden, the oldest average age, and a predominance of men, consistent with prior epidemiological evidence linking cardiovascular diseases disproportionately to older males (Benjamin et al., 2019; Roth et al., 2020). In contrast, the Depression+ cluster was more prevalent among women and strongly associated with mental health conditions, reflecting the well-established gendered patterns of mental health documented in previous research (McLean et al., 2011; Kuehner, 2017).

A further main contribution of this study is the integration of a relatively comprehensive set of social factors into the clustering framework. Using logistic regression, we linked social determinants—including demographic, socioeconomic indicators, behavioural factors, and psychosocial measures—to cluster membership, thereby illuminating the social heterogeneity underpinning multimorbidity. Behavioural factors such as alcohol consumption and physical activity emerged as key differentiators. While historical alcohol use was higher in clusters with greater disease severity, current alcohol use was lower, potentially reflecting behaviour modification following disease onset—an observation consistent with prior longitudinal studies (Shield et al., 2014). Physical health severity measures further reinforced the robustness of our clustering, with hypertension-related multimorbidity clusters consistently showing poorer health outcomes, higher medication use, and greater functional limitations. These findings align with literature on the compounded risks associated with multimorbidity involving cardiovascular and respiratory conditions (Violan et al., 2014; Cassell et al., 2018). Moreover, respiratory symptoms such as wheeze or chest whistling were disproportionately represented in asthma-related clusters, corroborating clinical heterogeneity observed in prior population-based studies (see To et al. (2012)).

In summary, our findings suggest that combining comprehensive disease data, advanced non-linear clustering techniques, and a broad set of social determinants can provide useful insights into the complexity of multimorbidity. This integrated approach helps to identify meaningful clusters while also highlighting the social heterogeneity underlying multimorbidity risk. While our work points to potential implications for public health policy —particularly the value of tailoring interventions to the characteristics of specific multimorbidity profiles—these findings should be interpreted with caution. Future research should focus on longitudinal analyses to track disease progression and clarify causal relationships within these multimorbid patterns. Incorporating a broader range of psychosocial and environmental factors will be essential to gain deeper insights into the etiology and progression of chronic diseases. Furthermore, as chronic conditions typically persist throughout an individual's life, investigating disease trajectories can provide valuable insights into effective disease management strategies, particularly in the context of rising multimorbidity prevalence in ageing populations.

Supplementary Information

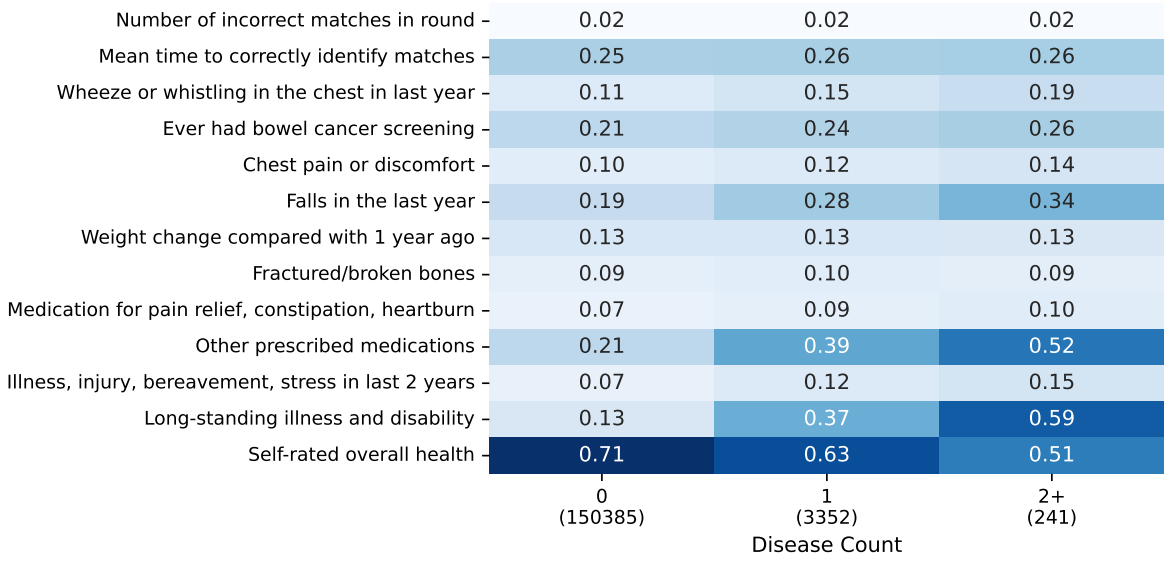

Figure S1: Physical severity within the healthy group by condition counts

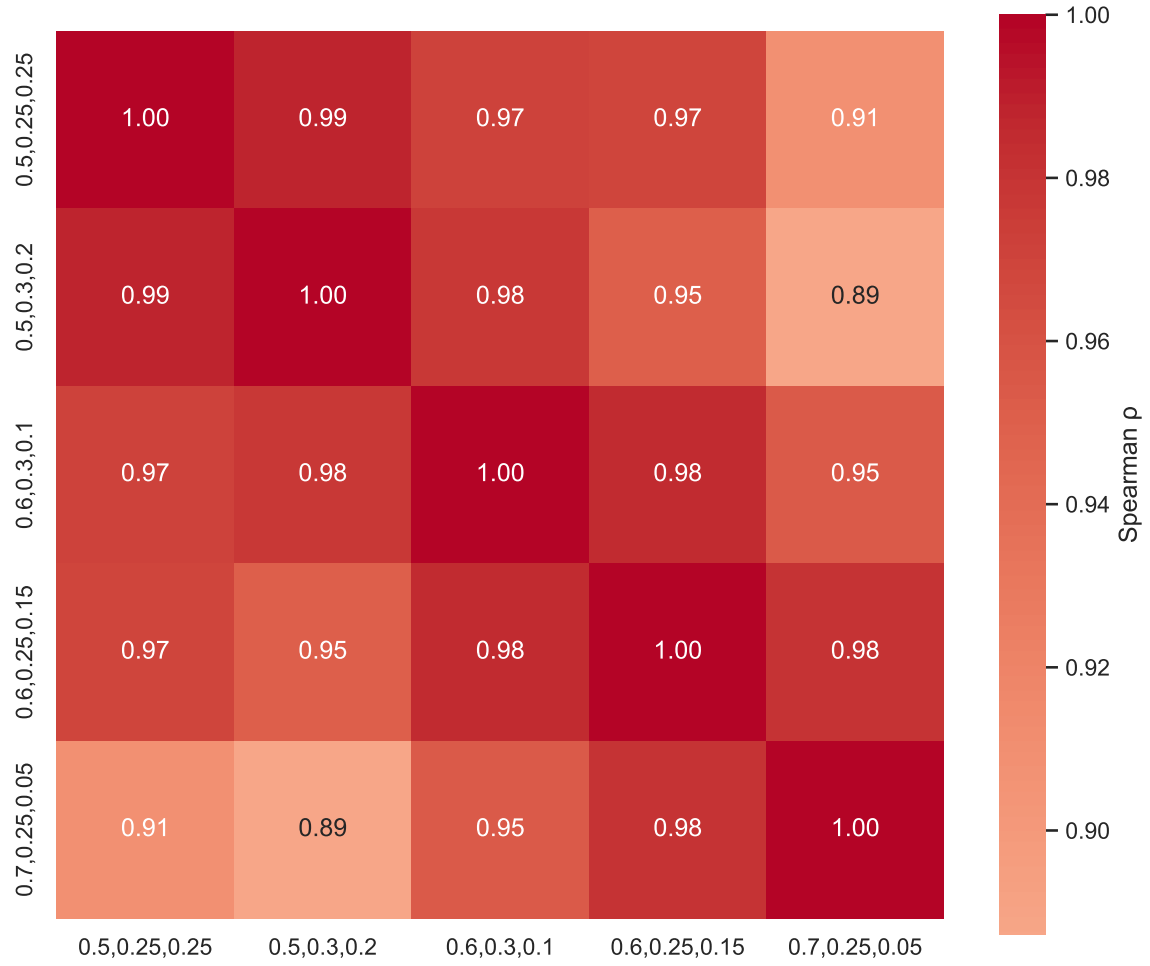

Figure S2: **Spearman ρ correlation across different weight configurations for $\mathscr{L}$.** This figure illustrates the correlations obtained under varying combinations of α, β, and γ (same order in the plot). In the main analysis, these parameters are set to 0.5, 0.25, and 0.25, respectively. Here, we compare alternative choices and show that the resulting correlations remain consistently high across weight configurations. This robustness suggests that our results are not sensitive to the specific choice of weights for $\mathscr{L}$.

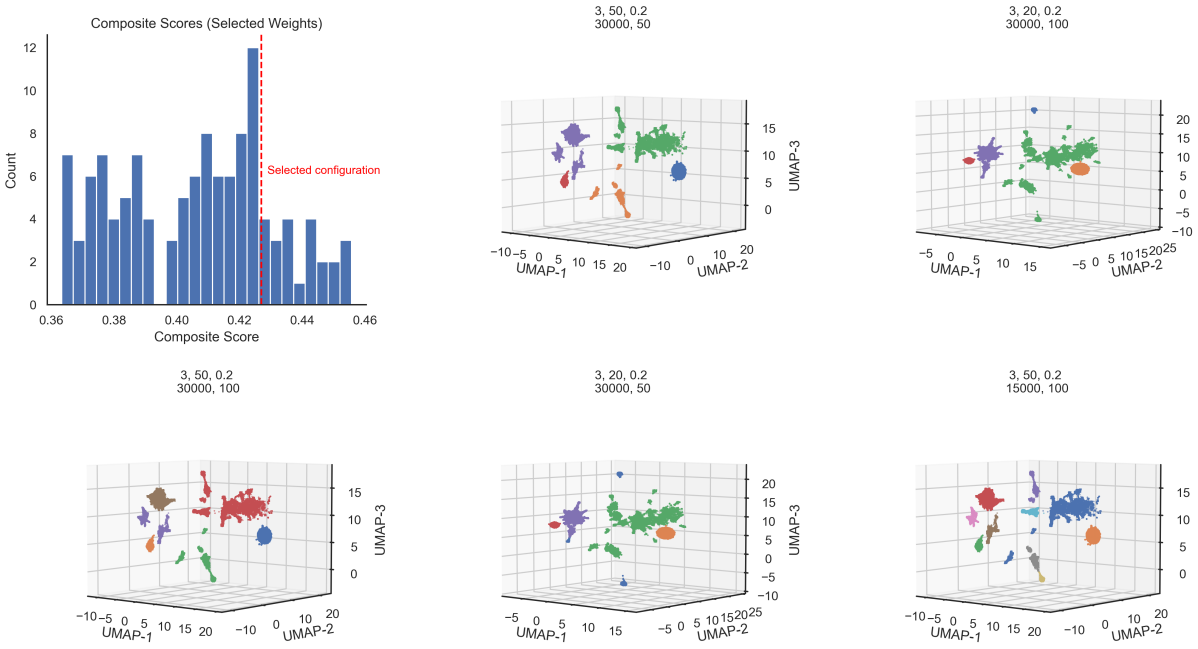

Figure S3: **Robustness of hyper-parameter selection.** The top left panel shows the distribution of the composite score $\mathscr{L}$ across different sets of hyperparameters within our grid search space, with each configuration repeated over five random seeds. The rest panels present the UMAP + HDBSCAN clustering results for the top five hyper-parameter sets ranked by $\mathscr{L}$. The results indicate that even though the absolute optimal configuration was not selected, the clustering structures are highly stable and separable across the top-performing parameter sets. Consequently, we prioritise the configuration that provides a finer-grained partition of the data (e.g., decomposing larger groups into smaller, more interpretable clusters), which allows us to better capture social heterogeneities and nuanced subgroup differences.

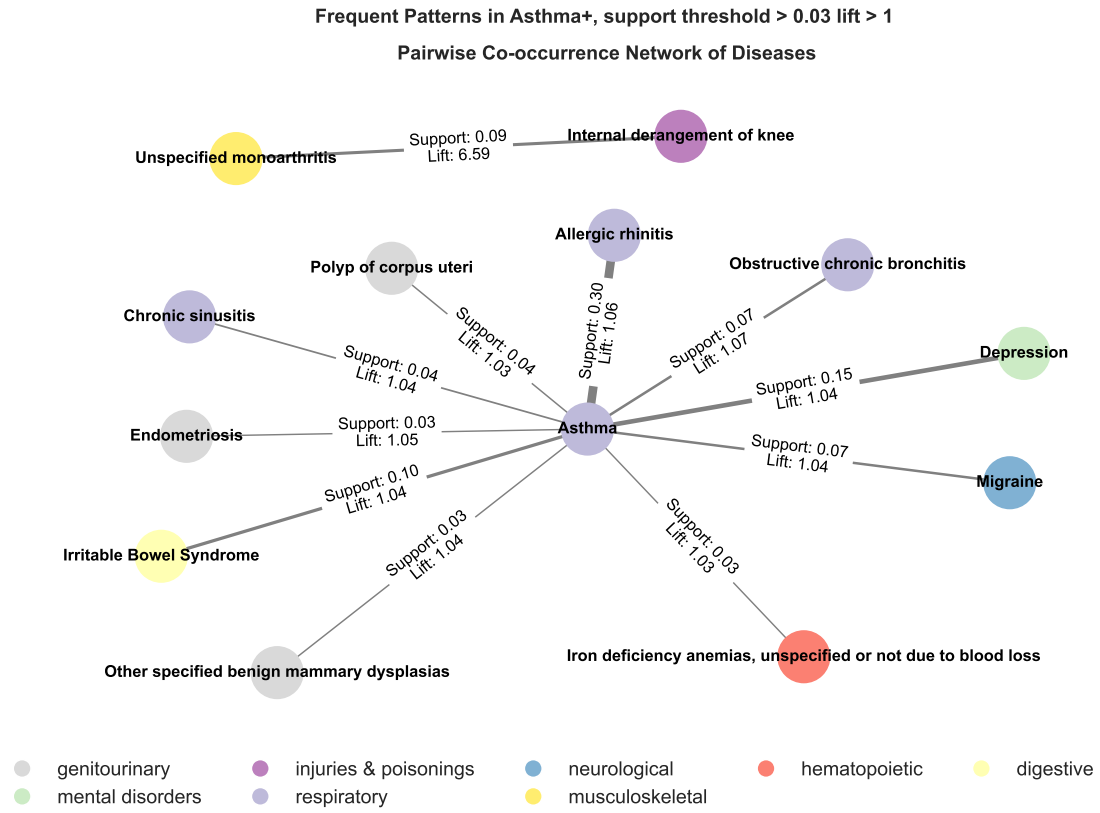

Figure S4: Network view of frequent co-occurrence patterns in Asthma+, where support threshold > 0.01 and lift > 1

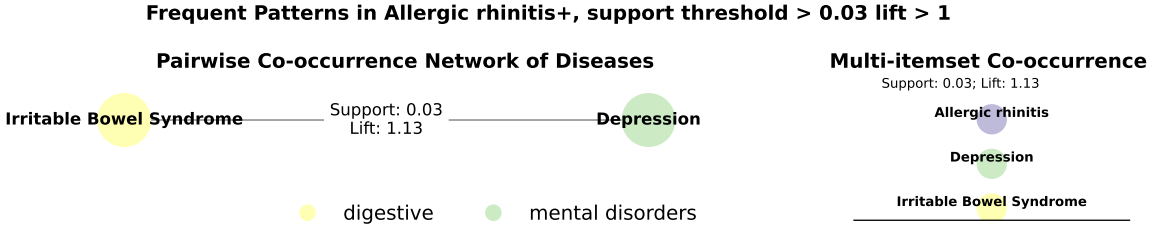

Figure S5: Network view of frequent co-occurrence patterns in Allergic rhinitis+, where support threshold > 0.03 and lift > 1

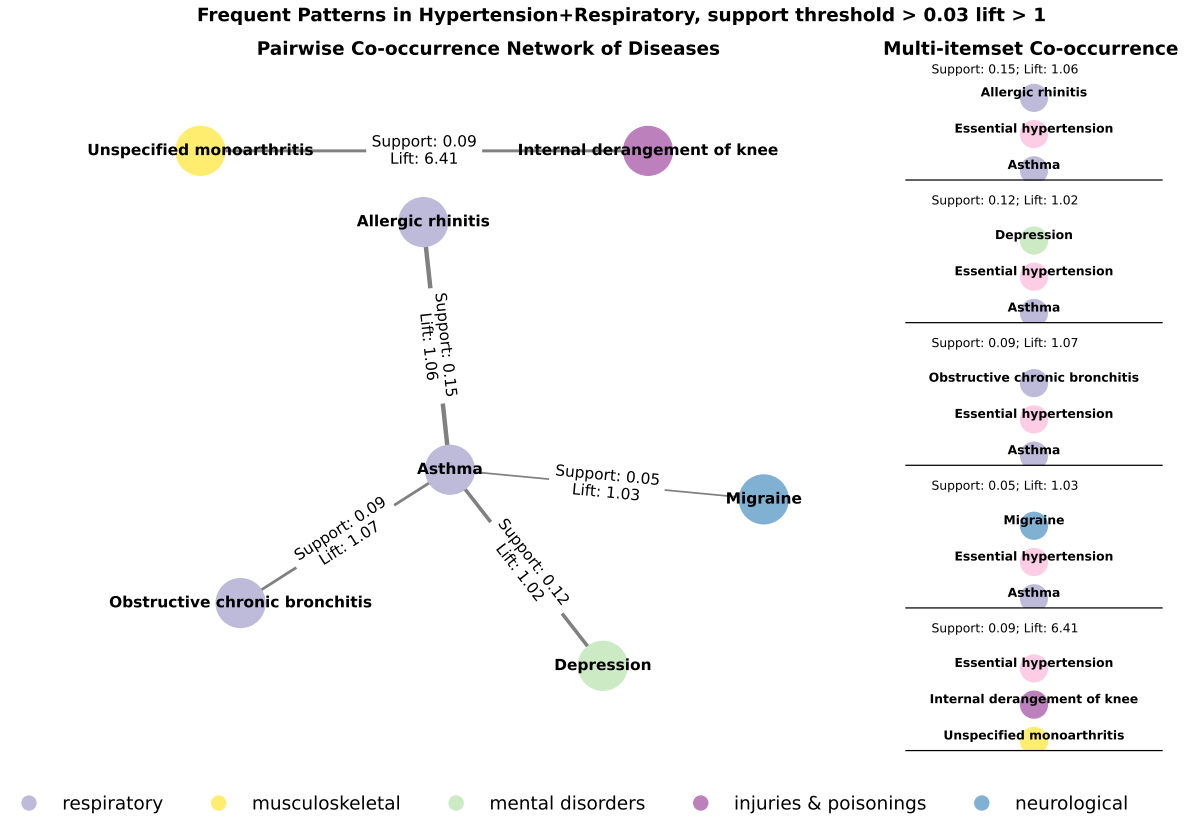

Figure S6: Network view of frequent co-occurrence patterns in Hypertension+Respiratory, where support threshold > 0.03 and lift > 1

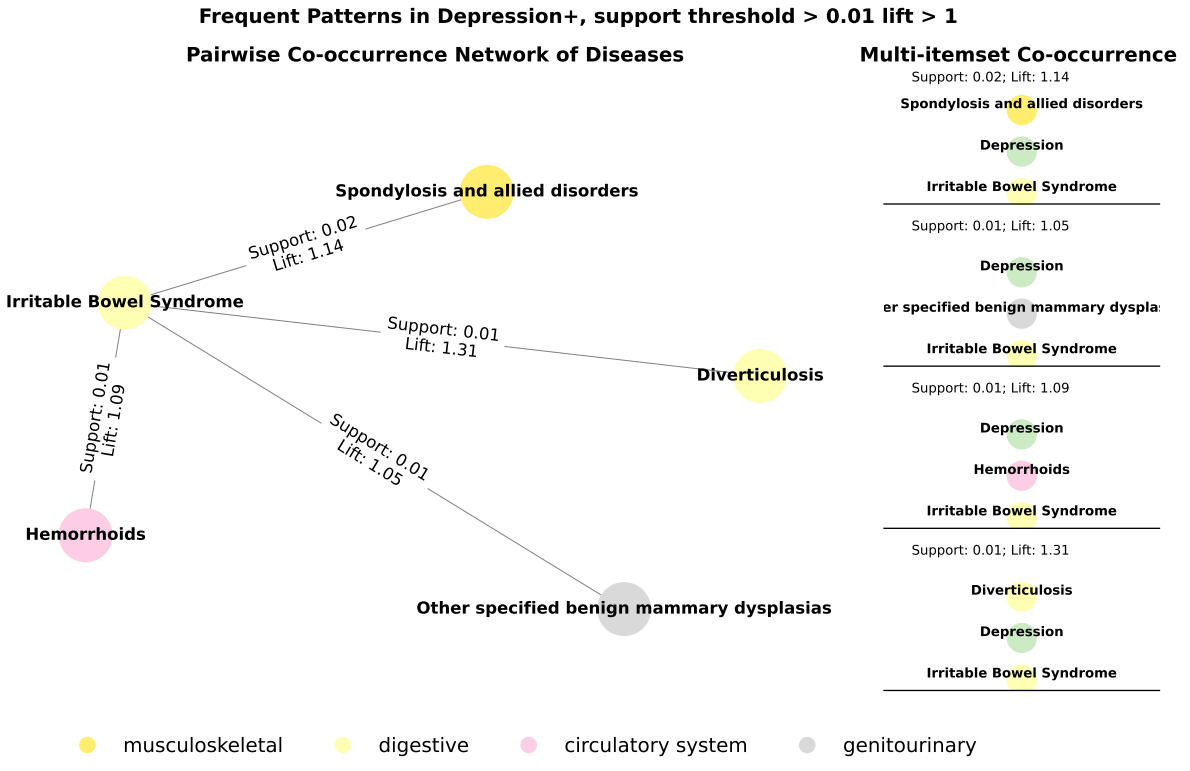

Figure S7: Network view of frequent co-occurrence patterns in Depression+, where support threshold > 0.01 and lift > 1

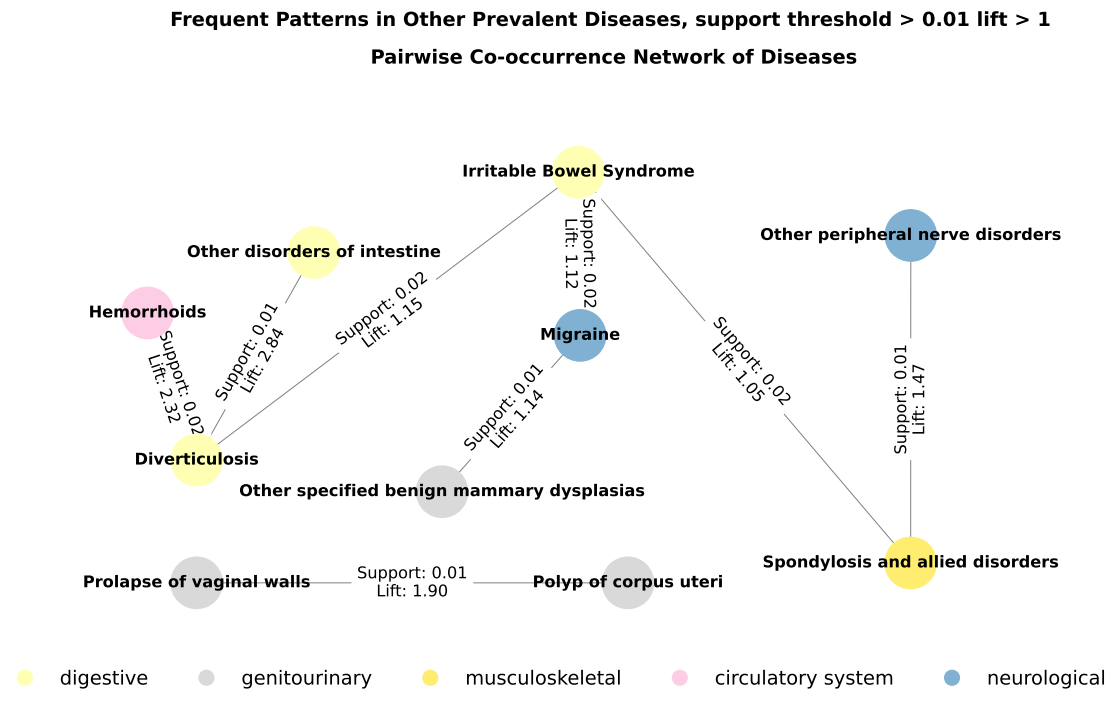

Figure S8: Network view of frequent co-occurrence patterns in Other Prevalent Diseases, where support threshold > 0.01 and lift > 1

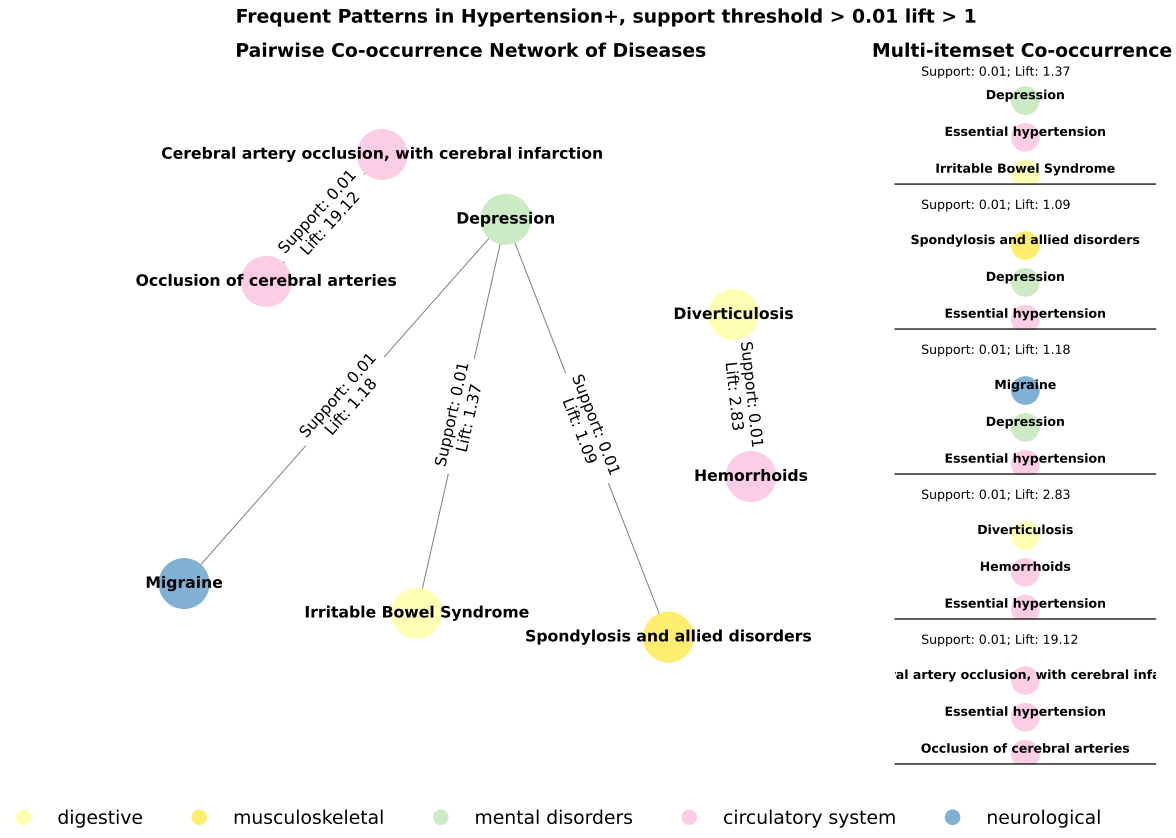

Figure S9: Network view of frequent co-occurrence patterns in Hypertension+, where support threshold > 0.01 and lift > 1

Phecode	Describe	ICD-10 codes
401.1	Essential hypertension	I10 Essential (primary) hypertension
495.0	Asthma	J45 Asthma; J45.0 Predominantly allergic asthma; J45.1 Nonallergic asthma; J45.8 Mixed asthma; J45.9 Asthma, unspecified
476.0	Allergic rhinitis	J30 Vasomotor and allergic rhinitis; J30.0 Vasomotor rhinitis; J30.1 Allergic rhinitis due to pollen; J30.2 Other seasonal allergic rhinitis; J30.3 Other allergic rhinitis; J30.4 Allergic rhinitis, unspecified
296.2	Depression	F32 Depressive episode; F32.0 Mild depressive episode; F32.1 Moderate depressive episode; F32.2 Severe depressive episode without psychotic symptoms; F32.3 Severe depressive episode with psychotic symptoms; F32.9 Depressive episode, unspecified
564.1	Irritable Bowel Syndrome	K58 Irritable bowel syndrome; K58.0 Irritable bowel syndrome with diarrhoea; K58.9 Irritable bowel syndrome without diarrhoea
340.0	Migraine	G43 Migraine; G43.0 Migraine without aura [common migraine]; G43.2 Status migrainosus; G43.8 Other migraine; G43.9 Migraine, unspecified; G44.0 Cluster headache syndrome
411.3	Angina pectoris	I20 Angina pectoris; I20.1 Angina pectoris with documented spasm; I20.8 Other forms of angina pectoris; I20.9 Angina pectoris, unspecified
721.0	Spondylosis & allied disorders	M47 Spondylosis; M47.9 Spondylosis, unspecified; M48.3 Traumatic spondylopathy
562.1	Diverticulosis	K57 Diverticular disease of intestine; K57.0 Diverticular disease of small intestine with perforation and abscess; K57.1 Diverticular disease of small intestine without perforation or abscess; K57.2 Diverticular disease of large intestine with perforation and abscess; K57.3 Diverticular disease of large intestine without perforation or abscess; K57.4 Diverticular disease of both small and large intestine with perforation and abscess; K57.5 Diverticular disease of both small and large intestine without perforation or abscess; K57.8 Diverticular disease of intestine, part unspecified, with perforation and abscess; K57.9 Diverticular disease of intestine, part unspecified, without perforation or abscess
411.8	Other chronic ischemic heart disease	I25 Chronic ischaemic heart disease; I25.5 Ischaemic cardiomyopathy; I25.6 Silent myocardial ischaemia; I25.8 Other forms of chronic ischaemic heart disease; I25.9 Chronic ischaemic heart disease, unspecified

Table S1: Mapping between the Phecodes and ICD-10 codes for the top 10 prevalent chronic diseases

References

Agency for Healthcare Research and Quality. (2024). Chronic Condition Indicator Refined (CCIR) for ICD-10-CM, v2024.1. Retrieved June 4, 2024, from https://www.hcup-us.ahrq.gov/toolssoftware/ccir/ccir.jsp

Álvarez-Gálvez, J., Ortega-Martín, E., Carretero-Bravo, J., Pérez-Muñoz, C., Suárez-Lledó, V., & Ramos-Fiol, B. (2023). Social determinants of multimorbidity patterns: A systematic review. *Frontiers in Public Health*, *11*, 1081518. https://doi.org/10.3389/fpubh.2023.1081518

Beaney, T., Clarke, J., Salman, D., Woodcock, T., Majeed, A., Aylin, P., & Barahona, M. (2024). Identifying multi-resolution clusters of diseases in ten million patients with multimorbidity in primary care in England [Publisher: Nature Publishing Group]. *Communications Medicine*, *4*(1), 102. https://doi.org/10.1038/s43856-024-00529-4

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2019). Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*, *139*(10), e56–e528. https://doi.org/10.1161/CIR.0000000000000659

Cassell, A., Edwards, D., Harshfield, A., Rhodes, K., Brimicombe, J., Payne, R., & Griffin, S. (2018). The epidemiology of multimorbidity in primary care: A retrospective cohort study [Publisher: British Journal of General Practice Section: Research]. *British Journal of General Practice*, *68*(669), e245–e251. https://doi.org/10.3399/bjgp18X695465

Dalmia, A., & Sia, S. (2021, December). Clustering with UMAP: Why and How Connectivity Matters [arXiv:2108.05525 [cs]]. https://doi.org/10.48550/arXiv.2108.05525

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

Jiang, Y., Zhao, B., Wang, X., Tang, B., Peng, H., Luo, Z., Shen, Y., Wang, Z., Jiang, Z., Wang, J., Ye, J., Wang, X., & Zhu, H. (2025). UKB-MDRMF: A multi-disease risk and multimorbidity framework based on UK biobank data [Publisher: Nature Publishing Group]. *Nature Communications*, *16*(1), 3767. https://doi.org/10.1038/s41467-025-58724-3

Kingston, A., Robinson, L., Booth, H., Knapp, M., Jagger, C., & for the MODEM project. (2018). Projections of multi-morbidity in the older population in England to 2035: Estimates from the Population Ageing and Care Simulation (PACSim) model. *Age and Ageing*, *47*(3), 374–380. https://doi.org/10.1093/ageing/afx201

Krauth, S. J., Steell, L., Ahmed, S., McIntosh, E., Dibben, G. O., Hanlon, P., Lewsey, J., Nicholl, B. I., McAllister, D. A., Smith, S. M., Evans, R., Ahmed, Z., Dean, S., Greaves, C., Barber, S., Doherty, P., Gardiner, N., Ibbotson, T., Jolly, K., … Jani, B. D. (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 [Publisher: Elsevier]. *eClinicalMedicine*, *74*. https://doi.org/10.1016/j.eclinm.2024.102703

Kuehner, C. (2017). Why is depression more common among women than among men? *The Lancet Psychiatry*, *4*(2), 146–158. https://doi.org/10.1016/S2215-0366(16)30263-2

Launders, N., Hayes, J. F., Price, G., & Osborn, D. P. (2022). Clustering of physical health multimorbidity in people with severe mental illness: An accumulated prevalence analysis of United Kingdom primary care data [Publisher: Public Library of Science]. *PLOS Medicine*, *19*(4), e1003976. https://doi.org/10.1371/journal.pmed.1003976

Marengoni, A., Rizzuto, D., Wang, H.-X., Winblad, B., & Fratiglioni, L. (2009). Patterns of Chronic Multimorbidity in the Elderly Population. *Journal of the American Geriatrics Society*, *57*(2), 225–230. https://doi.org/10.1111/j.1532-5415.2008.02109.x

McInnes, L., & Healy, J. (2017). Accelerated Hierarchical Density Clustering [arXiv:1705.07321 [stat]]. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42. https://doi.org/10.1109/ICDMW.2017.12

McInnes, L., Healy, J., & Melville, J. (2020, September). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [arXiv:1802.03426 [stat]]. https://doi.org/10.48550/arXiv.1802.03426

McLean, C. P., Asnaani, A., Litz, B. T., & Hofmann, S. G. (2011). Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *Journal of Psychiatric Research*, *45*(8), 1027–1035. https://doi.org/10.1016/j.jpsychires.2011.03.006

Ng, S. K., Tawiah, R., Sawyer, M., & Scuffham, P. (2018). Patterns of multimorbid health conditions: A systematic review of analytical methods and comparison analysis. *International journal of epidemiology*, *47*(5), 1687–1704. https://doi.org/10.1093/ije/dyy134

Nichols, L., Taverner, T., Crowe, F., Richardson, S., Yau, C., Kiddle, S., Kirk, P., Barrett, J., Nirantharakumar, K., Griffin, S., Edwards, D., & Marshall, T. (2022). In simulated data and health records, latent class analysis was the optimum multimorbidity clustering algorithm. *Journal of Clinical Epidemiology*, *152*, 164–175. https://doi.org/10.1016/j.jclinepi.2022.10.011

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., … Fuster, V. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*, *76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010

Shield, K. D., Parry, C., & Rehm, J. (2014). Chronic Diseases and Conditions Related to Alcohol Use. *Alcohol Research : Current Reviews*, *35*(2), 155–171. Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908707/

Stafford, M., Steventon, A., Thorlby, R., Fisher, R., & Deeny, S. (2018). *Briefing: Understanding the health care needs of people with multiple health conditions*. Health Foundation London.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

To, T., Stanojevic, S., Moores, G., Gershon, A. S., Bateman, E. D., Cruz, A. A., & Boulet, L.-P. (2012). Global asthma prevalence in adults: Findings from the cross-sectional world health survey. *BMC Public Health*, *12*(1), 204. https://doi.org/10.1186/1471-2458-12-204

Violan, C., Foguet-Boreu, Q., Flores-Mateo, G., Salisbury, C., Blom, J., Freitag, M., Glynn, L., Muth, C., & Valderas, J. M. (2014). Prevalence, Determinants and Patterns of Multi-

morbidity in Primary Care: A Systematic Review of Observational Studies [Publisher: Public Library of Science]. *PLOS ONE*, *9*(7), e102149. https://doi.org/10.1371/journal.pone.0102149

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J. C., Theodoratou, E., & Wei, W.-Q. (2019). Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR medical informatics*, *7*(4), e14325. https://doi.org/10.2196/14325

Zemedikun, D. T., Gray, L. J., Khunti, K., Davies, M. J., & Dhalwani, N. N. (2018). Patterns of Multimorbidity in Middle-Aged and Older Adults: An Analysis of the UK Biobank Data. *Mayo Clinic Proceedings*, *93*(7), 857–866. https://doi.org/10.1016/j.mayocp.2018.02.012