Understanding Symptom Patterns in COVID-19 Using Dimensionality Reduction and Clustering Methods

A extended research project report submitted to the University of Manchester for the degree of master of Data Science in the Faculty of Humanities

Student ID: 11558899

Year of submission: 2025

The Name of School: School of Social Science

Contents

Word count:

List of Illustrations

Abstract

Acknowledgements

I want to sincerely thank everyone who has helped and supported me over my academic path as I finish my postgraduate thesis.

First and foremost,I want to pay my respects to my supervisor,Prof.House.Thank you for your expertise,patient guidance and valuable advice.Your academic rigor and passion for research not only inspired me but also provided a clear direction for my academic exploration.

My classmates and friends,your help and support have also been an important factor in enabling me to complete my dissertation.The time we spent together,whether it was hard studying in the library or relaxing time on campus,will be my precious memories.Once again,I want to express my gratitude in particular to my housemates,who have supported me greatly in my academic endeavors in addition to being wonderful friends throughout my life.Your understanding,encouragement and occasional academic discussions during the dissertation writing process have greatly helped me to maintain a clear mind and a positive attitude.

I also want to express my gratitude to my family,especially to my parents.Their love and support has been an inexhaustible source of motivation for me to keep moving forward.You have always given me the greatest understanding and encouragement in my pursuit of my academic dreams.

DECLARATION

This dissertation submitted to The University of Manchester is my original work unless referenced clearly to the contrary, and no portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

INTELLECTUAL PROPERTY STATEMENT

i. The author of this extended research project report (including any appendices and/or schedules to this report) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this report, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the report, for example graphs and tables ("Reproductions"), which may be described in this report, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this report, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant dissertation restriction declarations deposited in the University Library, The University Library's regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University's Guidance for the Presentation of dissertations.

1. Introduction

Symptoms are basic descriptive concepts in clinical practice and disease studies that provide clues to disease processes. In practice, pure single symptoms are hard to find. Different types of symptoms tend to co-occur in individuals and form co-occurrence patterns. With the rapid development of large-scale health data and computational methods, symptom co-occurrence analysis has emerged as an important interdisciplinary topic in epidemiology, clinical medicine and data science.

The usefulness of symptom co-occurrence analysis has been demonstrated in infectious and chronic diseases. During the pandemic of coronavirus disease 2019 (COVID-19), community surveys and clinical cohort studies reported different types of symptom clusters including various symptom combinations ranging from respiratory patterns to digestive patterns or neurological patterns. Different symptom clusters are associated with different disease severities (Menni et al., 2020; Liu et al., 2021). Similar clustering phenomena were also observed in three chronic diseases: fibromyalgia, depression and rheumatoid arthritis. Symptom co-occurrence is associated with disease classfication and prognosis (Wolfe et al., 2018). Discovering structured relationships between symptoms provides a new way to capture disease diversity and provide evidences for diagnosis and treatment.

Although the potential of analysing symptom co-occurrence exists, methodological challenges still exist. Symptom data can be represented as a binary matrix. As the number of individuals increases and the number of symptom dimensions increases, the data tends to be high-dimensional and sparse. Traditional statistical methods are hard to extract structural information from such data. Many studies adopt longitudinal designs and follow up the same individuals in multiple occasions. However, it further increases challenges to analyse such data. It has been observed that the prevalence of different symptoms and clustering patterns are stage-specific with respect to diseases. For example, during the acute stage, respiratory symptoms are more common while during the recovery stage, fatigue or neurological symptoms are more common. In chronic diseases, different symptom combinations also exist in different stages. Therefore, methods need

to capture temporal changes without fitting noise. The data we analysed in this study has both binary and longitudinal characteristics and that is why methods need to adopt dimension reduction and clustering methods that can reduce sparsity, capture dynamics and be robust to noise.

Dimension reduction provides an alternative way to analyse symptom co-occurrence. The core idea of dimension reduction is to project high-dimensional data into a low-dimensional space while retaining structural information in high-dimensional data. This makes analysis easier and interpretation clearer. Principal Component Analysis (PCA) is the most classical method. It achieves this by finding linear directions that capture maximal variance. However, PCA has limitations in representing binary structures and non-linear structures (Landgraf & Lee, 2015). Non-negative matrix factorisation (NMF) provides another way to factorise sparse non-negative data. Its decomposition results in interpretable components that can be connected to symptom groups including respiratory system or digestive system clusters (Lee & Seung, 1999; Brunet et al., 2004). This makes NMF attractive to be clinically interpreted.

In recent years, non-linear methods have begun to enter this space. t-SNE has shown considerable promise at visualising high-dimensional data, preserving local neighbourhood information and revealing latent clusters (van der Maaten & Hinton, 2008) but performs poorly at preserving global relationships and at very large sample sizes. UMAP is based on similar ideas but unites manifold learning with topological theory to preserve local and global information better while being computationally more efficient (McInnes et al., 2018). UMAP performs well in single-cell transcriptomics (Becht et al., 2019) and is increasingly being applied to clinical symptom datasets to reveal clusters under binary and longitudinal contexts. When applied to clinical symptom data, UMAP can be applied to longitudinal symptom data to capture the longitudinal trajectories of a patient's symptom evolution.

Similarity measures play a central role in co-occurrence analysis. As symptoms can be thought of as sets, the Jaccard coefficient is the most commonly used similarity measure to calculate the degree to which two sets co-occur by dividing their intersection by their union (Levandowsky & Winter, 1971). Similarity matrices derived from the Jaccard coefficient are often used as input to clustering analysis. For instance, hierarchical clustering produces clustering diagrams where

multi-level structures are visualised by progressively merging similar symptoms or individuals together (Murtagh & Contreras, 2012). Previous work has shown that applying Jaccard similarity in combination with visualisation through UMAP can generate interpretable and clinically meaningful symptom clusters (Elliott et al., 2023). In longitudinal studies, these clusters can be extended to different time points to allow researchers to visualise changes in cluster membership and co-occurrence patterns across the disease course.

Ethical and privacy considerations should be prioritised throughout the entire process of analysing symptom co-occurrence. As symptom data forms a type of sensitive health data, strict ethical and legal guidelines should be followed in its use. The UK Government's Data Ethics Framework highlights transparent, fair and accountable use of data (Xafis, Vicki, et al., 2019) and the General Data Protection Regulation (GDPR) sets out principles of minimising data use, lawfulness and being traceable (Voigt & Von dem Bussche, 2017). This supports the creation of a TRE, a secure environment in which sensitive health data can be accessed. Additionally, the widespread use of machine learning in health research has raised ethical concerns around algorithmic bias, interpretability and trustworthiness (Floridi & Cowls, 2019). Therefore, for binary and longitudinal symptom data, research methods not only need to be accurate in their results but also ethically correct in how we use data.

Although there has been some work in applying dimensionality reduction and clustering methods to health data, very few studies have systematically explored the ability of these methods to apply to symptom co-occurrence data in binary form and longitudinal form. Most existing research focus on static binary data, few methods can handle both features. To bridge this gap, this study evaluates a set of complementary techniques including principal component analysis (PCA), non-negative matrix factorisation (NMF), t-SNE, UMAP and clustering methods based on Jaccard similarity. The study evaluates these methods using both simulated and real world datasets to see whether we can robustly discover underlying symptom clusters, capture their dynamic changes over time and discover interpretable information useful for clinical and public health practice.

2. Literature Review

Symptom co-occurrence research lies at the intersection of epidemiology, clinical medicine, and data science. Compared to existing methods (which only studied single symptom or limited clinical indicators), analysis of co-occurrence matrix offers a new way to explore the patterns of symptom combinations. These new patterns not only help to describe disease phenotypes, but also offer a methodological basis to explore individual differences, study disease progression and inform personalized interventions. Previous studies have explored this topic from the following aspects: the impact of symptom co-occurrence on disease heterogeneity, application of dimension reduction method, selection of similarity measure and clustering method, construction of ethical framework for health data research.

Infectious disease research provides a unique use case for studying symptom co-occurrence. Research has shown that infected individuals show multiple representative clusters of symptoms rather than a single typical presentation. Menni et al. (2020) applied a large-scale mobile app survey to cluster patients into groups based on different symptom patterns and found that different symptom patterns are associated with disease severity and recovery outcomes. In the study of elderly patients in Wuhan, researchers found a link between symptom co-occurrence and cognitive decline (Liu et al., 2021). When studying the complexity of long COVID, researchers stressed the importance of symptom clustering as a critical step when exploring the post-infection syndrome (Aiyegbusi et al., 2021). These studies show the potential of using comorbidity analysis to study infectious diseases and the potential for further applications to chronic diseases. In fibromyalgia, Wolfe et al (2018) modified the diagnostic criteria based on symptom combinations and argued that classification should focus on symptom combinations rather than single symptoms. These findings indicate that comorbidity analysis can offer consistent methodological and empirical support across different diseases.

When data is highly co-occuring and sparse, dimension reduction is needed. Principal Component Analysis (Collins et al., 2002) is used due to its simplicity, but it is linearly assumed which hinders its applicability on binary symptom data. For this reason Logical PCA was developed to improve the capturing of binary matrices. Dimension reduction is achieved by projecting data onto the natural parameter space (Landgraf and Lee, 2015). This method performs very well when applied

on psychological questionnaires, genetic , or clinical data (Collins et al., 2002), but computational costs prohibit its application on populations.

Non-negative matrix factorisation (Lee and Seung, 1999; Brunet et al., 2004) provides an alternative approach to dimensionality reduction for sparse non-negative data. Non-negative matrix factorisation can output components that can be interpreted as symptom clusters, for example respiratory or digestive system clusters. More recent approaches, such as t-SNE and UMAP, build on these ideas for non-linear structures. t-SNE preserves the local neighbourhoods of the data and shows the inferred latent clusters, however it does this at the expense of global relationships (van der Maaten and Hinton, 2008). UMAP is built on manifold learning with topological concepts and is computationally efficient whilst maintaining both local and global features (McInnes et al., 2018). This method has been applied widely in single-cell genomics (Becht et al., 2019) and is being applied increasingly to analyse symptom data, aiding in the detection of phenotypic clustering within cross-sectional and longitudinal data.

Similarity measures can also be used to identify symptom co-occurrence. As symptoms can be considered as sets, the Jaccard matrix is especially suitable for binary data. It calculates the ratio of shared symptoms to total symptoms and is less biased towards the Euclidean distance measure in sparse binary structures (Levandowsky and Winter, 1971). Elliott et al. (2023) applied Jaccard-based similarity to SARS-CoV-2 community infections and showed multiple phenotypic clusters. Hierarchical clustering is often applied alongside the Jaccard matrix, and the clustering graph structure of this can reveal nested relationships between symptom groups (Murtagh and Contreras, 2012). Related methods have also been explored combining model-based clustering and hierarchical methods which can ensure statistical inference and robustness (Fraley and Raftery, 2002; Kriegel et al., 2009).

In addition to algorithms, symptom research should also address ethical and governance issues around symptom data. Symptom data is sensitive data that should be handled according to the law. The UK Data Ethics Framework principles state that data researchers should be transparent, accountable and fair (UK Government, 2020). The Alan Turing Institute (2019) state that the risks associated with data use must be assessed and biases removed at each stage of the research process.

From a broader perspective, the EU's General Data Protection Regulation (GDPR) sets out legal and technical standards for the use of personal data (Voigt and Von dem Bussche, 2017) which have led to the adoption of requirements such as secure and controlled analysis provided by Trusted Research Environments (TRE) (Garcia et al., 2018). The widespread use of machine learning and artificial intelligence has also led to concerns around data ethics, bias and public data trust. Floridi and Cowls (2019) state that results must be both scientifically sound and socially acceptable.

This chapter discusses the necessity of using dimension reduction and clustering methods to analyse symptom co-occurrence phenomena, while acknowledging the technical and ethical challenges involved. Disease heterogeneity demonstrates the need to consider symptom patterns and different methods such as PCA, NMF, t-SNE and UMAP have advantages for different data characteristics. Different similarity measures can be used to interpret the Jaccard matrix, including hierarchical clustering. Ethical and governance frameworks ensure that this analysis is responsible and trustworthy. Building on this fundation, this study seeks to evaluate the performance of a set of dimensionality reduction and clustering methods on both simulated and real-world datasets. The aim is to assess whether they can provide stable symptom clusters in both binary and longitudinal data and also to explore their relevance for future research in clinical medicine and public health.

3. Methodology

3.1 Research Design

This study evaluates the feasibility,robustness,and generalisability of dimensionality reduction and clustering methods in the analysis of symptom co-occurrence.Symptom data in epidemiology are often high-dimensional,sparse,and longitudinal,which limits the effectiveness of traditional statistical tools.The question is whether dimensionality reduction can uncover latent symptom clusters across different environments and support the interpretation of symptom phenotypes.

The research is organised into three stages.First,simulated trajectory data are used to test if methods can recover pre-defined cluster structures.This validates their ability under controlled dynamic conditions and reflects longitudinal features.Second,sparse binary symptom matrices are constructed to mimic real epidemiological surveys,where most entries are zeros and only a few symptoms co-occur.This tests robustness in high-dimensional sparse settings.Third,the methods are applied to real-world Jaccard similarity matrices from three UK COVID-19 monitoring systems:the Infection Survey(CIS),the Second Generation Surveillance System(SGSS),and Pillar 2 community testing.Consistency across these datasets indicates generalisability;differences may reveal the influence of data collection or population structure.

This layered design builds a logical chain:trajectory data validate methods,binary matrices test sparse environments,and real data confirm robustness across populations.The study asks three central questions:can latent structures be identified,can effectiveness be maintained with sparse binary data,and can results remain consistent across real-world datasets?Progressing from idealised to real settings ensures methodological clarity while preserving practical relevance.

3.2 The Generation of Simulated Trajectory Dataset and Preprocessing

The generation of trajectories is based on a random walk model with drift.Two-dimensional random walk trajectory data is designed and generated,and then expanded into three-dimensional space through non-linear projection.In this experiment,each individual is represented as a random trajectory generated on a two-dimensional plane,and the category of the trajectory is determined by different drift vectors.In this way,potential cluster structures can be artificially introduced to test whether dimension reduction and clustering methods can recover predefined category relationships without supervised information.The generation of trajectories is based on a random walk model with drift.Let the position of the i-th trajectory at time step j be $X_{i,j} \in \mathbb{R}^2$ Xi,j ∈ R2,then the recursive relationship is:

$$X_{i,j} = X_{i,j-1} + J_{c(i)} + \varepsilon_{i,j}, \quad j = 1, \ldots, T-1, \tag{1}$$

where is the drift vector determined by the latent class c(i),and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2 I_2)$ represents Gaussian noise.This model ensures that trajectories belonging to different classes exhibit distinct overall trends,while local noise introduces overlap,thereby mimicking the structure often observed in heterogeneous epidemiological data.

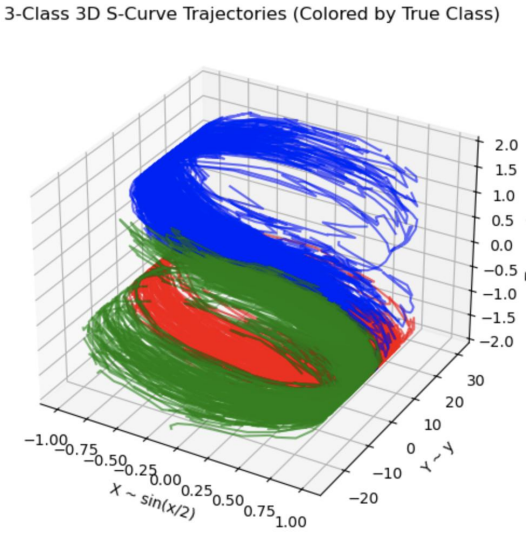

Figure 1: 3D Trajectories colored by true class

To enhance geometric complexity,the two-dimensional trajectories were mapped into a three-dimensional space via the nonlinear transformation:

$$Z_{i,j} = f(X_{i,j}) = \left(\sin\left(\frac{x_{i,j}}{2}\right),\ 2y_{i,j},\ \text{sign}(x_{i,j}) \cdot \left(\cos\left(\frac{x_{i,j}}{2}\right) - 1\right) \right). \qquad (2)$$

This transformation preserves the time order and introduces bending and folding characteristics,simulating the nonlinear relationships found in real data,so that category distinctions no longer rely on simple linear boundaries.

In order to further reduce the impact of trajectory starting points and overall shifts on subsequent analysis,this study did not directly use three-dimensional position sequences,but instead performed time differences on them:

$$\Delta Z_{i,j} = Z_{i,j} - Z_{i,j-1}, \quad j = 1, \ldots, T-1, \tag{3}$$

This approach focuses on local dynamic patterns rather than absolute positions,which is more in line with the focus of symptom co-occurrence analysis.

Since the differenced components may take positive or negative values,they were decomposed into complementary nonnegative parts to enable compatibility with algorithms such as Nonnegative Matrix Factorization(NMF):

$$(\Delta Z_{i,j})^{+} = \max(\Delta Z_{i,j}, 0), \quad (\Delta Z_{i,j})^{-} = \max(-\Delta Z_{i,j}, 0). \tag{4}$$

Each trajectory was then represented by concatenating the vectorized positive and negative components across all time steps:

$$\phi_i = \left[\mathrm{vec}((\Delta Z_{i,\cdot})^{+}),\ \mathrm{vec}((\Delta Z_{i,\cdot})^{-}) \right] \in \mathbb{R}^{2\cdot(T-1)\cdot 3}. \tag{5}$$

Aggregating across all individuals yielded the feature matrix $\Phi \in \mathbb{R}^{N\times 2\cdot(T-1)\cdot 3}$,which served as the standardized input for subsequent dimensionality-reduction and clustering analyses.

The simulated data generation process includes four key steps:two-dimensional random walks,three-dimensional nonlinear projections,time difference processing,and positive and negative decomposition.This design retains the preset category labels for easy comparison and verification,while adding complexity through noise and nonlinearity to provide uniform and realistic input conditions for methodological experiments.It is the core starting point of the research framework.

3.3 The Generation of Simulated Binary Dataset

In the second type of experiment,a simulated binary symptom dataset was generated to mimic the common“presence or absence”format of symptoms in epidemiological and clinical research.The

dataset is represented as a matrix $X \in \{0,1\}^{N \times M}$,where N denotes the number of simulated individuals and M the number of symptoms.Each element is defined as x_{ij}=1 if individual i exhibits symptom j,and x_{ij}=0 otherwise.This binary structure reflects the form of symptom data frequently collected through surveys and electronic health records.

Unlike the dynamic random-walk trajectories,the binary symptom matrix focuses on modeling latent group differences.Several clusters were defined,each characterized by distinct symptom occurrence probabilities.For an individual iii in cluster k,symptom j was sampled according to:

$$x_{ij} \sim \text{Binomial}(1, p_{kj}), \quad i \in \text{cluster } k, \tag{6}$$

where pjk is the probability that symptom j occurs within cluster k.With n=1,this reduces to a Bernoulli distribution,though implemented via a binomial form to facilitate batch generation.

In practice,clusters were assigned markedly different probability profiles.For example,in one cluster,some symptoms had low probabilities (e.g.,p=0.08,0.10,0.12 p=0.08,0.10,0.12 p=0.08,0.10,0.12), others moderate values (e.g.,p=0.27,0.30,0.37 p=0.27,0.30,0.37 p=0.27,0.30,0.37), while a subset reached high probabilities(e.g.,p=0.66,0.75,0.82,0.91 p=0.66,0.75,0.82,0.91 p=0.66,0.75,0.82,0.91). Another cluster concentrated its high-probability symptoms in different dimensions, while a third cluster exhibited generally low probabilities.

This design reproduces the mixed structure of“common”and“cluster-specific”symptoms observed in real diseases.For instance,one cluster might display high probabilities for fever and cough,resembling a respiratory pattern,while another emphasizes dyspnea and chest pain,reflecting alternative phenotypes.The simulated data thus retain sparsity while embedding latent heterogeneity,enabling dimensionality-reduction and clustering methods to capture meaningful group structures.

3.4 Real-World Jaccard Datasets and Analytical Workflow

To complement the simulation experiments,real-world surveillance data were introduced to assess the robustness and generalizability of dimensionality reduction and clustering methods.Data came from three major UK COVID-19 monitoring systems:the COVID-19 Infection Survey(CIS),the Second Generation Surveillance System(SGSS),and the Pillar 2 community testing programme.These sources collectively cover community surveys,laboratory-confirmed clinical cases,and large-scale community testing,offering complementary perspectives on symptom co-occurrence.

Instead of using individual-level records,the study relied on pre-computed Jaccard similarity matrices,defined as:

$$J(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \tag{7}$$

where S_i and S_j denote the sets of individuals reporting symptoms i and j.The Jaccard index ranges from 0(no overlap) to 1(perfect co-occurrence), capturing structural relationships while avoiding privacy concerns.

Each dataset provides distinct advantages.CIS,managed by the Office for National Statistics,systematically sampled households and longitudinally followed participants,thus reflecting community-level symptom co-occurrence,including mild or asymptomatic cases.SGSS,based on statutory laboratory reporting,captures clinically tested individuals and reveals patterns linked to confirmed infections.Pillar 2,with fixed testing sites and home kits,offers wide coverage and large sample sizes,enabling detection of population-level co-occurrence patterns.SGSS and Pillar 2 used identical symptom definitions,supporting direct comparison,while CIS included fewer items,offering a test of generalizability across heterogeneous symptom spaces(Fyles et al., 2023).

For analysis,UMAP was applied for nonlinear dimensionality reduction,hierarchical clustering was conducted using $1-J$ as a distance metric,and heatmaps with dendrogram-based ordering visualized similarity strength.Together,these methods provided systematic comparison across datasets,evaluating both consistent and dataset-specific clustering patterns.

3.5 Experimental Methods and Algorithms

As introduced in the previous section,the input for the real-data experiments in this study is the Jaccard similarity matrix calculated based on symptom co-occurrence relationships,whose formal definition is provided in Section 3.4.Building on this,this section will introduce various dimension reduction and clustering methods applied to simulated trajectory data and simulated binary symptom data to evaluate the robustness and generalisation capability of the algorithms in capturing potential symptom phenotypes.The methods employed include traditional linear dimensionality reduction(PCA),non-negative matrix factorisation(NMF),as well as recently popular non-linear manifold learning methods(t-SNE,UMAP).These are complemented by partition-based and hierarchical clustering methods(k-means,hierarchical clustering)and intuitive heatmap analysis.

3.5.1 Principal Component Analysis(PCA)

PCA is a classical linear dimensionality reduction method that seeks directions of maximum variance in the data(Jolliffe,2002).Given a data matrix $X \in \mathbb{R}^{n\times p}$ with covariance matrix$\Sigma = \frac{1}{n}X^\top X$,the optimization objective is

$$\max_{u} \; u^\top \Sigma u \quad \text{s.t. } \|u\| = 1. \tag{8}$$

Its solutions are the eigenvectors of Σ ,called principal components,with eigenvalues reflecting explained variance.Selecting the top k components yields $Z=XU_k$.In this study,PCA served as a baseline:while the Jaccard similarity matrix is not strictly Euclidean,PCA provides a useful linear benchmark against which nonlinear methods can be compared.

3.5.2 Non-negative Matrix Factorization(NMF)

NMF is a dimensionality reduction method widely applied in text mining and bioinformatics,particularly suited for nonnegative data.Its basic formulation is

$$X \approx WH, \quad W, H \geq 0, \tag{9}$$

where $X \in \mathbb{R}_{+}^{n \times p}$ is the input matrix, $W \in \mathbb{R}_{+}^{n \times k}$ represents individuals in a reduced latent space,and $H \in \mathbb{R}_{+}^{k \times p}$ captures the contribution of symptoms to latent components(Lee&Seung,1999).Unlike PCA,NMF ensures nonnegativity,yielding interpretable“parts-based”components,e.g.,a respiratory cluster or a gastrointestinal cluster.In this study,NMF was applied to both simulated trajectory data and simulated binary data to test whether stable latent symptom patterns could be extracted.

3.5.3 k-means Clustering

k-means is one of the most widely used partitioning clustering methods,aiming to minimize the within-cluster sum of squared errors(SSE):

$$\min_{C} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2, \tag{10}$$

where μ_k denotes the centroid of cluster k(MacQueen,1967).Its main advantage is computational efficiency,producing rapid clustering results.However,it is sensitive to initialization and requires the number of clusters K to be specified in advance.In this study,k-means was applied to identify potential symptom clusters,which were then compared with hierarchical clustering results.

3.5.4 Hierarchical Clustering

Hierarchical clustering builds a tree structure by iteratively merging clusters,thus revealing aggregation patterns across multiple scales(Sokal&Michener,1958). Here,agglomerative clustering was employed,starting from individual symptoms and progressively merging until a complete hierarchy was formed.The distance metric was defined as the complement of the Jaccard similarity:

$$d(i,j) = 1 - J(i,j). \tag{11}$$

Average linkage was chosen as the aggregation criterion to reduce sensitivity to extreme values.The resulting dendrograms illustrate the hierarchical structure of symptoms,helping to interpret which clusters may correspond to latent phenotypic categories.

3.5.5 t-distributed Stochastic Neighbor Embedding(t-SNE)

t-SNE is a nonlinear dimensionality reduction method widely used for high-dimensional data visualization,aiming to preserve local neighborhood structure through probability distributions(van der Maaten&Hinton,2008).In the high-dimensional space,the similarity between points iii and j is defined as:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \tag{12}$$

while in the low-dimensional space a corresponding distribution q_{ij} is defined using a Student-t kernel:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}. \tag{13}$$

The optimization objective minimizes the Kullback–Leibler divergence:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{14}$$

t-SNE effectively reveals local cluster structures but is less reliable at preserving global topology.In this study,it was used mainly as a comparison to UMAP for exploring alternative visualizations.

3.5.6 Uniform Manifold Approximation and Projection(UMAP)

UMAP is a manifold learning–based method for nonlinear dimensionality reduction(McInnes et al.,2018).It first constructs a k-nearest-neighbor graph in the high-dimensional space,transforms it into a weighted fuzzy simplicial set,and then optimizes a cross-entropy loss to preserve local neighborhoods in the low-dimensional embedding.The objective is:

$$C = \sum_{(i,j)} w_{ij} \log \frac{w_{ij}}{\hat{w}_{ij}}, \tag{15}$$

where w_{ij} are high-dimensional edge weights and $\hat{w}_{ij}$ the corresponding low-dimensional probabilities.UMAP balances local and global structure preservation while being computationally

efficient and scalable.In this study,it was the primary nonlinear method applied to Jaccard matrices,serving both for visual exploration of symptom clusters and comparison with PCA and t-SNE.

3.5.7 Heatmaps and Visualisation Analysis

In addition to dimensionality reduction and clustering,this study also comprehensively utilises various visualisation techniques to more intuitively present the relationships between symptoms.Specifically,the low-dimensional embedding plots of UMAP,t-SNE,and PCA demonstrate the distribution characteristics and potential cluster structures of symptoms in low-dimensional space;the results of hierarchical clustering are presented through dendrograms to illustrate the hierarchical relationships between different symptoms;Additionally,the Jaccard similarity matrix is output as a heatmap,with colour intensity reflecting the co-occurrence strength between symptoms,enabling researchers to grasp the overall similarity patterns across different datasets.These visualisation methods complement each other:low-dimensional embeddings aid in identifying potential cluster structures,hierarchical clustering reveals hierarchical relationships,while heatmaps provide a global perspective.

3.5.8 Robustness and Generalizability Assessment

A stepwise validation framework was designed to evaluate methodological robustness and generalizability.First,simulated trajectory data were used to test whether each method could recover pre-defined cluster structures under controlled conditions.Second,simulated binary symptom data allowed assessment of performance in sparse co-occurrence environments resembling survey data.Finally,real Jaccard similarity matrices were employed to verify stability and consistency in large-scale surveillance datasets.By applying a unified analytical pipeline across these three settings,the study ensured comparability between synthetic and real-world results,providing stronger evidence for the validity of dimensionality reduction and clustering methods in symptom co-occurrence research.

4. Results and Analysis

4.1 Overview

After completing the method design,this chapter will focus on presenting and analysing experimental results under different datasets and algorithm conditions.The study involves three types of data:first,trajectory dataset used to simulate longitudinal data;second,binary dataset used to reproduce symptom co-occurrence patterns in clinical and epidemiological surveys;third,real-world data from the UK's CIS,SGSS,and Pillar 2 systems,with input in the form of Jaccard similarity matrices,reflecting symptom structures across different populations and channels.

In terms of methodology,the experiments employed PCA, NMF, t-SNE, UMAP, k-means,and hierarchical clustering,supplemented by heatmap visualisation for cross-validation of results.The results will cover the recovery of cluster structures in trajectory data using different dimensionality reduction and clustering methods,the decomposition effect of NMF on potential symptom clusters in binary data,and the differences and commonalities in embedding space and clustering structures among the three monitoring systems in real-world data.Additionally,this chapter evaluates the robustness and generalisability of the methods through cross-dataset comparisons and explores the reproducibility of symptom patterns between simulated and real-world data.

4.2 Trajectory Data Experiment Results and Analysis

The first step of this study was to evaluate the dimension reduction and clustering methods used based on three types of trajectory data systematically.The data was generated using random walks with drift and Gaussian noise,and projected into a three-dimensional space to form‘S’-shaped curves,exhibiting distinct nonlinearity and inter-class overlap characteristics.

4.2.1 Clustering Results

This experiment first compared the performance of K-means,hierarchical clustering,and NMF.K-means outperformed the other methods overall.Its confusion matrix showed that the

third-class trajectories could be aggregated relatively accurately,but there was still significant misclassification between the first and second classes.The Normalised Mutual Information(NMI)was 0.6027,indicating that the method could capture the global cluster structure,but its accuracy decreased when the boundaries were affected by noise.

Hierarchical clustering performed relatively poorly.Although it can gradually build hierarchical relationships,under conditions of highly non-linear and overlapping trajectories,the clustering boundaries shift significantly,with a large number of third-category samples incorrectly assigned to the second category,causing the NMI to drop to 0.4406.This indicates that the method is sensitive to noise and struggles to reliably distinguish complex cluster structures.

NMF offers another approach based on feature decomposition.Its non-negative decomposition partially extracts the latent patterns of trajectories,but the overall accuracy is limited,with an NMI of only 0.4801.While there is some improvement in the identification of the third category,confusion between the first and second categories remains significant,indicating that it fails to fully address the issue of nonlinear overlap.

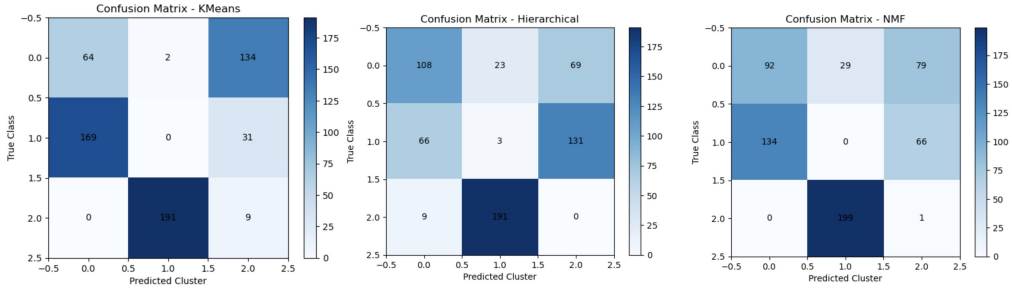

Figure 2: Confusion matrix for three cluster methods

4.2.2 Dimension Reduction Results

In the dimension reduction analysis, PCA, t-SNE, and UMAP demonstrate notably different behaviors (see Figure 3). PCA preserves the linear trend and global variance structure of the data to some extent. However, its performance is limited when dealing with complex nonlinear trajectories, as evidenced by the partial overlap between the first and second categories along the decision boundaries.

t-SNE shows strong local clustering capabilities, with certain trajectory groups being clearly distinguished in the two-dimensional embedding. Nonetheless, it distorts the global data structure to a degree, which can make inter-cluster relationships difficult to interpret and reduce its usefulness in capturing broader data geometry.

UMAP produces the most informative embedding. It maintains the overall S-shaped trajectory distribution while also achieving relatively clean cluster separation, successfully balancing global topology and local neighborhood preservation. This supports UMAP's strength in handling nonlinear and high-dimensional data.

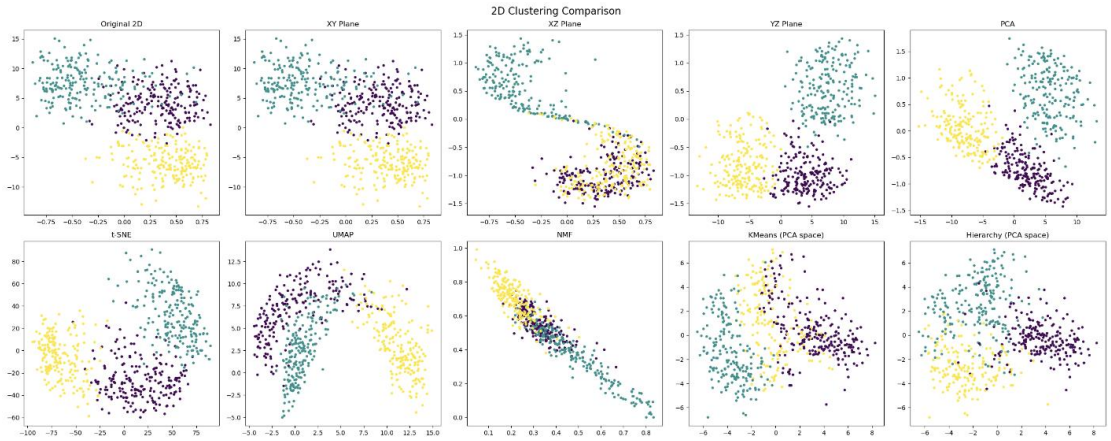

Figure 3 :Comparison of 2D projections

4.2.3 Comprehensive Discussion

A comprehensive analysis of clustering and dimensionality reduction results shows that traditional clustering algorithms are limited in their performance on complex trajectory data.K-means slightly outperforms others in overall clustering accuracy but exhibits significant boundary misclassification;hierarchical clustering performs even worse and struggles to handle high-noise environments;NMF provides some improvement in feature decomposition but does not lead to significant overall enhancement.

The comparison of dimension reduction methods highlights the importance of structure preservation.PCA is stable but lacks nonlinear expressive capability;t-SNE has advantages in local discrimination but sacrifices global interpretability;UMAP strikes a balance between the

two,preserving overall structure while achieving high resolution locally,providing more reliable low-dimensional input for subsequent clustering and phenotype identification.

In summary,the trajectory experiments indicate that relying solely on traditional clustering is insufficient for achieving optimal results when handling complex dynamic data.Combining advanced non-linear dimensionality reduction methods(especially UMAP)significantly enhances clustering accuracy and interpretability.

4.3 Binary Data Experiment Results and Analysis

In the second set of experiments,we evaluated various dimension reduction and clustering methods using a synthetic binary symptom dataset.Unlike the trajectory experiments,each dimension of this dataset is a 0/1 binary variable representing whether a particular symptom is present.By setting different symptom occurrence probabilities across clusters,we simulated the symptom co-occurrence patterns commonly observed in clinical and epidemiological studies.

4.3.1 Clustering Results

The two-dimensional visualisation of K-means clustering applied to the binary data (see Figure 4) shows three clearly separated clusters, broadly consistent with the simulated class structure. The cluster boundaries are well-defined, indicating that the binary features retained sufficient separability in the low-dimensional space for effective clustering.

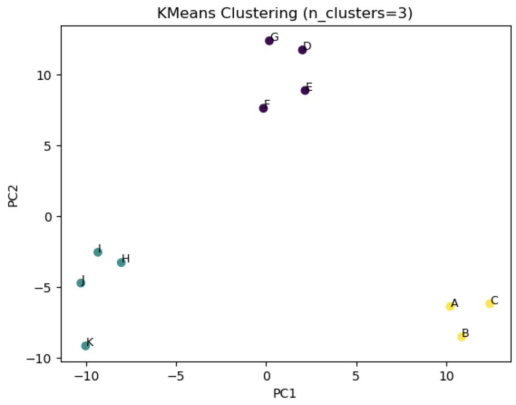

Figure 4: 2D visualisation of K-means clustering on binary data

The results of hierarchical clustering are presented via heatmaps and dendrograms(see Figure 5).It can be observed that some symptoms are merged earlier in the dendrogram,indicating higher co-occurrence probabilities,while others only aggregate at higher levels,reflecting their distinctiveness.This hierarchical information is valuable for interpreting relationships between symptoms,but the overall clustering clarity is inferior to K-means.

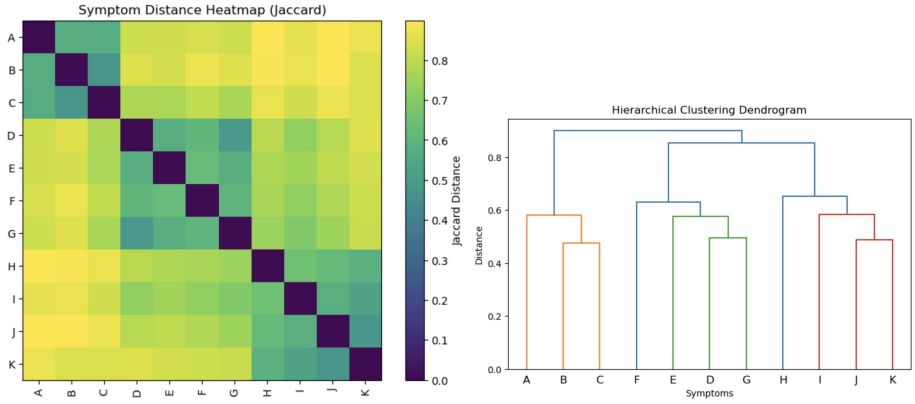

Figure 5:Heatmap and dendrogram of hierarchical clustering on binary data

The NMF results offer an interpretable factor-based representation of the binary data. By decomposing the symptom matrix into non-negative latent components, NMF captures continuous variations in symptom expression across individuals. While the resulting factor space does not form clearly separated clusters, it reveals underlying gradients of symptom co-occurrence, offering supplementary insight into the structural composition of symptom patterns.

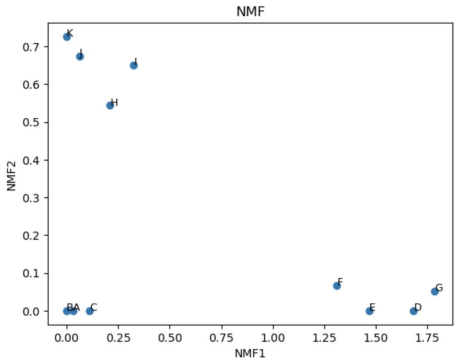

Figure 6: 2D visualisation of NMF on binary data

4.3.2 Dimension Reduction Results

In the comparison of dimension reduction methods,PCA,t-SNE and UMAP exhibit significant differences in performance.The PCA projection(see Figure 7 left) preserves a clear global structure, with samples forming three distinct clusters corresponding to the original grouping. Despite the linear nature of PCA, the binary symptom data retains sufficient variance to achieve separability in the first two principal components.

The t-SNE projection (Figure 7, middle) results in a nearly uniform and isotropic distribution of samples, with no evident clustering or meaningful structure. This indicates that, under the current parameter setting (perplexity=10), t-SNE fails to capture any underlying symptom co-occurrence patterns in the binary data.

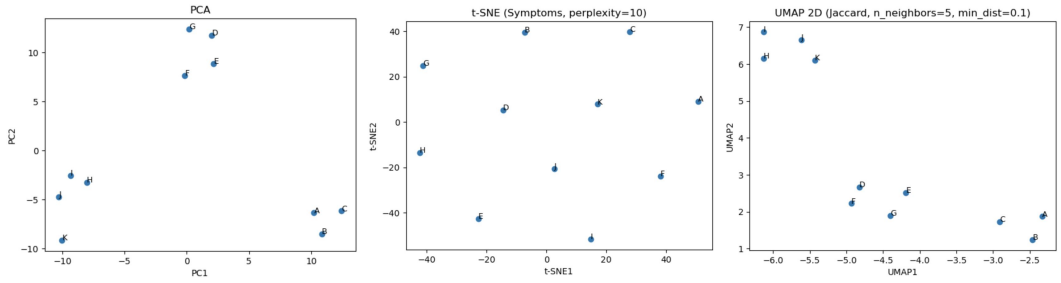

Figure 7:Comparison of two-dimensional projections of PCA,t-SNE,and UMAP

UMAP's results are the most prominent(see Figure 7 right),and it is particularly sensitive to parameter selection in binary data.This experiment also examines the performance of UMAP under varying hyperparameters. When n_neighbors is small (e.g., 3), the model emphasizes local relationships, resulting in compact and well-separated clusters. In contrast, setting n_neighbors to 20 produces a more uniform and dispersed layout, where the original cluster structure is no longer visible. Adjusting min_dist also affects the compactness of the embedding: smaller values (e.g., 0.001) produce tighter groupings, while moderate values (e.g., 0.2) preserve clearer separation between clusters. These results confirm that UMAP is highly sensitive to parameter settings, and different configurations lead to different trade-offs between local cohesion and global structure. Therefore, careful parameter tuning is essential when applying UMAP to binary symptom data.

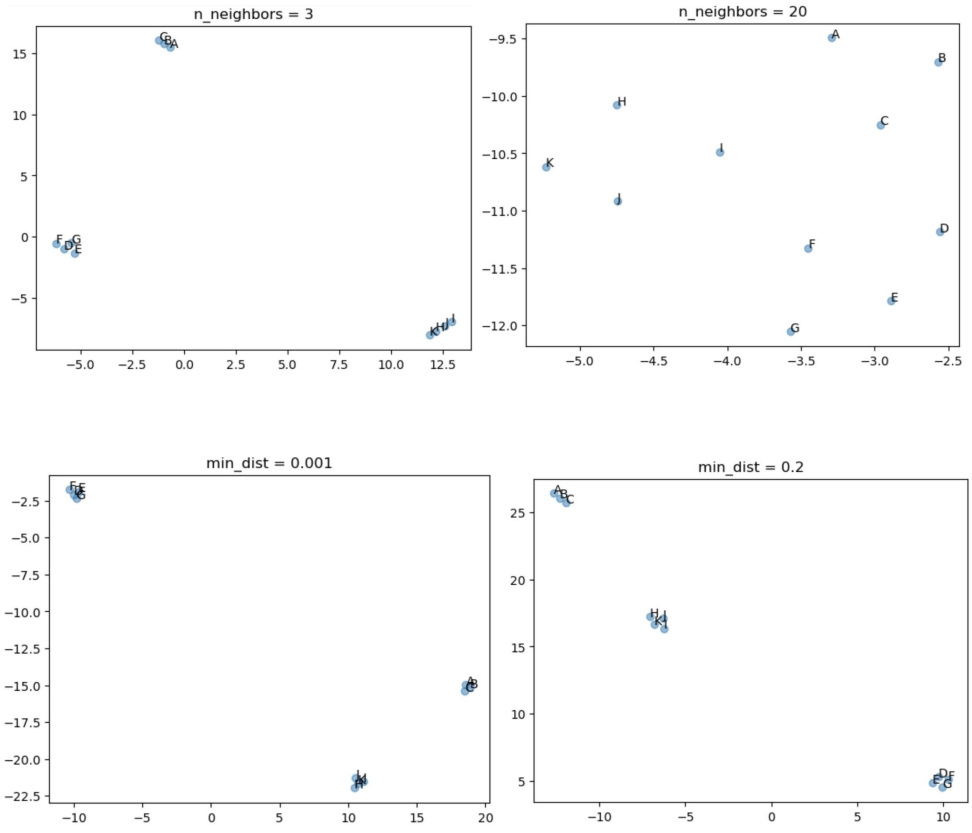

Figure 8:Comparison of two-dimensional projections of UMAP under different n_neighbors and min_dist parameters

4.3.3 Comprehensive Discussion

The binary data experiments revealed different pattern characteristics compared to the trajectory experiments.In linear dimensionality reduction and traditional clustering methods,the sparsity and non-linear relationships of binary data lead to insufficient separability:t-SNE struggles to form clear clusters,and hierarchical clustering is sensitive to noise.In contrast,PCA and UMAP perform better in local pattern recognition,with UMAP demonstrating high flexibility through parameter adjustment.NMF further provides factor-level explanations,aiding in the identification of potential symptom combination patterns.

4.4 Dimension Reduction Results and Analysis of the Jaccard Matrix Based on Real Datasets

This section will analyse the dimension reduction results of the Jaccard matrix based on real COVID-19 symptom datasets,with data sources including three UK surveillance systems:SGSS(Second Generation Surveillance System),Pillar 2 community testing,and CIS(COVID-19 Infection Survey).This experiment combines the results of UMAP and hierarchical clustering with heatmaps to analyse symptom co-occurrence patterns across different populations and data collection methods.

4.4.1 Dimension Reduction Results

UMAP projections reveal meaningful patterns of symptom co-occurrence across datasets(see Figure 9). In both SGSS and Pillar 2, respiratory symptoms (e.g., cough, sore throat) and systemic symptoms (e.g., fever, fatigue) form tight clusters, suggesting they represent core phenotypes of COVID-19. Altered state symptoms appear isolated at the periphery, reflecting their low frequency and weaker associations. In contrast, the CIS dataset shows a more dispersed structure, with symptoms like loss of smell and loss of taste appearing in distinct locations. This separation implies that these symptoms were not consistently co-reported in community-based surveys, potentially due to variation in symptom perception, recall bias, or questionnaire design. These findings highlight UMAP's ability to uncover latent symptom structures and dataset-specific reporting patterns.

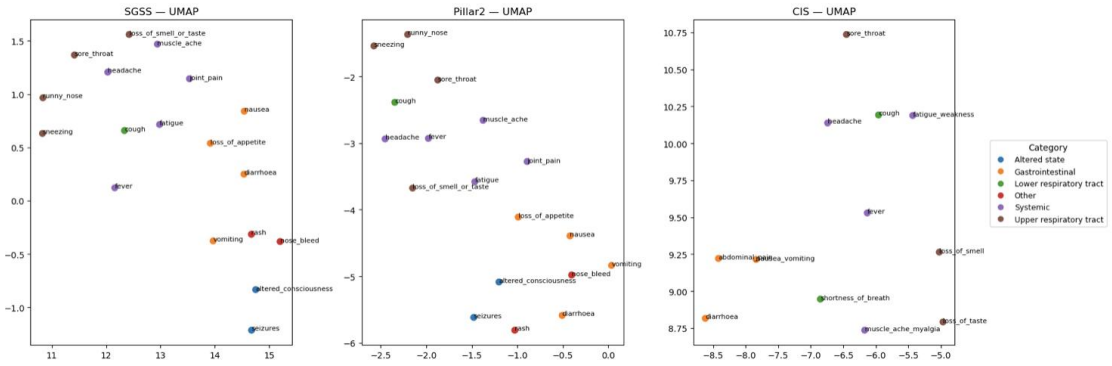

Figure 9:UMAP embedding results for the three datasets

4.4.2 Clustering Results

Hierarchical clustering based on the Jaccard coefficient reveals nested relationships among symptoms across the three datasets (see Figure 10). In SGSS and Pillar 2, respiratory symptoms such as cough, sore throat, and runny nose cluster tightly, while systemic symptoms like fever, fatigue, and headache form a separate but nearby group. These two groups eventually merge at higher linkage levels, suggesting a consistent pattern of symptom co-occurrence in structured testing settings. In contrast, CIS shows a more fragmented structure, with loss of smell and loss of taste clustering together but remaining distant from respiratory symptoms, reflecting greater variability in self-reported data. Rare symptoms such as rash and seizures consistently appear at the periphery of the dendrograms, indicating weak associations with core symptom clusters.

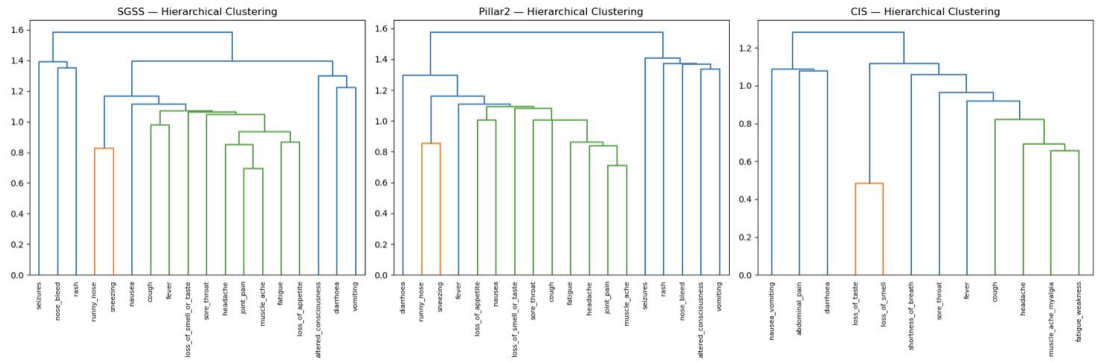

Figure 10:Hierarchical clustering dendrogram of the three datasets

4.4.3 Heatmap analysis

The heatmap visually displays pairwise symptom similarity based on Jaccard coefficients (Figure 11). In SGSS and Pillar 2, symptoms such as cough and sore throat, fever and headache show relatively high similarity, suggesting they form core co-occurrence patterns. The CIS dataset presents greater heterogeneity, with symptom similarities varying more widely. Notably, loss of smell and loss of taste remain highly correlated across all datasets. Meanwhile, altered state symptoms such as seizures and altered consciousness consistently exhibit low similarity with other symptoms, confirming their peripheral role in the overall pattern.

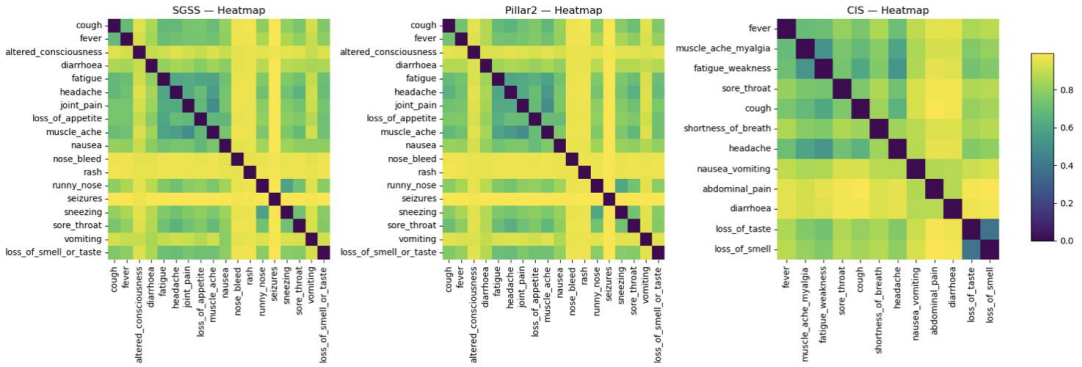

Figure 11:Jaccard similarity heatmap of the three datasets

4.4.4 Comprehensive Discussion

Combining the three types of data,respiratory and systemic symptoms consistently form the core cluster,reflecting the common clinical features of COVID-19.Differences primarily manifest in CIS,whose typical symptoms are more easily separated into independent units,while SGSS and Pillar 2 exhibit greater consistency in overall structure.Rare symptoms remain marginal across all datasets and are unlikely to constitute primary phenotypes.From a methodological perspective,UMAP,hierarchical clustering,and heatmaps complement each other:UMAP preserves global and local relationships,hierarchical clustering reveals hierarchical structures,and heatmaps visually present the co-occurrence intensity of symptom pairs.The combination of these three methods indicates that the approach used in this study demonstrates good stability in real-world monitoring data and possesses a certain degree of cross-dataset generalisation capability.

4.5 Discussion and Analysis

This study systematically examined dimension reduction and clustering methods in symptom co-occurrence analysis through three types of experiments:trajectory simulation,binary symptom simulation,and real monitoring data.The three types of experiments each served complementary roles in the research design:trajectory experiments simulated longitudinal processes to verify whether the methods could capture dynamic structures;binary data experiments highlighted sparse binary features to assess the algorithms'applicability in high-dimensional discrete spaces;and real-world data experiments tested their robustness under complex population and multi-data

source conditions.While these experiments built upon one another,they also formed a comparable control framework.

First,researchers can directly compare the consistency between the dimensionality reduction and clustering results in trajectory experiments and the actual categories.The results show that UMAP can simultaneously maintain global trajectory patterns and local separation,while t-SNE focuses more on local aggregation,and PCA tends to follow overall trends but loses its ability to distinguish non-linear patterns.In contrast,the binary data experiments introduce significant sparsity and heterogeneity.In this set of experiments,PCA's weaknesses are further amplified,while UMAP continues to demonstrate good discrimination performance in low-dimensional space and offers flexible trade-offs between local compactness and global balance through parameter tuning.NMF's advantages are more prominent in binary data,as its decomposition results directly correspond to potential symptom clusters,providing a structured foundation for interpreting sparse co-occurrence patterns.It can be seen that the two types of simulation experiments complement each other in terms of data characteristics:trajectories emphasise dynamic structures,while binary data emphasises sparse combinations.In both cases,UMAP and NMF outperform traditional methods,highlighting the necessity of non-linear dimensionality reduction and interpretable decomposition.

In the real datasets SGSS and Pillar2,respiratory and systemic symptoms consistently form core clusters,confirming the conclusion from the simulation experiments that‘major cluster patterns can be reliably identified.’In the CIS dataset,loss of smell and taste are separated independently,indicating that in broader,more heterogeneous population samples,the independence of typical symptoms may be highlighted.This contrasts with the findings from the trajectory and binary experiments:trajectory data emphasises the method's sensitivity to the overall pattern,binary data demonstrates its ability to capture sparse patterns,while real-world data reveals how these two capabilities migrate and transform across different population scenarios.

This progressive and contrasting relationship indicates that a single experiment is insufficient to support a comprehensive evaluation of method performance.Only by combining the three types of experiments can the effectiveness and robustness of the method be tested layer by layer,from dynamic simulation to sparse simulation to real-world monitoring data.Trajectory experiments demonstrate the method's potential in longitudinal structures,binary experiments highlight the

necessity of non-linear methods in sparse environments,and real-world data validate whether these advantages can be retained and generalised in actual populations.Together,they form a iterative validation framework:ideal environments confirm feasibility,sparse environments test robustness,and real-world environments validate generalisability.

The results of this study show that the combination of NMF and UMAP performs exceptionally well across all three types of data.NMF provides interpretable factor decomposition at the symptom level,while UMAP preserves nonlinear relationships in low-dimensional representations.Together,they form a complementary analytical framework that is both interpretable and stable.Jaccard similarity serves as a consistency metric,ensuring comparability across experiments and cross-dataset comparative analysis.Although real longitudinal symptom data has not yet been directly incorporated,the design of trajectory experiments lays the methodological groundwork for this direction.

5. Conclusion

The primary objective of this study is to evaluate and validate dimension reduction and clustering methods for symptom co-occurrence data, with the broader goal of revealing potential symptom phenotypes and developing analytical tools applicable to clinical research and public health surveillance. During the COVID-19 pandemic, identifying symptom co-occurrence patterns has become particularly important: such patterns reveal clinical heterogeneity, support case detection and disease surveillance, and provide a basis for policy responses. However, the core methodological challenge lies in extracting meaningful structures from binary and longitudinal data that are simultaneously high-dimensional, sparse, and noisy. To address this issue, we designed three complementary experimental settings—trajectory simulation, binary symptom simulation, and analysis of real-world Jaccard similarity matrices. These methods collectively provide methodological validation and empirical evidence, thereby combining theoretical exploration with application relevance.

Trajectory simulation aims to introduce temporal dynamics and serve as a simulation of longitudinal symptom data. By generating two-dimensional random walks and projecting them

into three-dimensional space, trajectories that evolve over time are constructed. The continuity and sequential nature of these trajectories parallel the temporal characteristics of longitudinal data, making them suitable as test cases for evaluating the performance of dimension reduction and clustering methods in dynamic environments. The results indicate that UMAP and t-SNE effectively preserve local neighbourhood relationships and distinguish trajectory groups in low-dimensional embeddings, while PCA, although capable of capturing overall trends, performs poorly in maintaining local structure. This suggests that non-linear methods are more suitable for extracting meaningful patterns from dynamic data and provides methodological guidance for the application of dimensionality reduction techniques in longitudinal symptom studies.

Binary symptom simulation aims to reproduce real clinical scenarios as realistically as possible. By setting symptom occurrence probabilities for different symptom groups, a sparse binary matrix is generated, whose structure is similar to symptom data in surveys or electronic health records. Such data is both sparse and contains unique symptom combinations, placing high demands on the robustness and interpretability of analytical methods. NMF performs exceptionally well in this context, as its non-negative decomposition directly extracts clinically meaningful symptom clusters, such as respiratory or digestive system groups. . UMAP embedding achieves clear separation of groups in a two-dimensional space, providing an intuitive supplement to the results. The combination of the two methods not only generates structured, interpretable factors but also presents intuitive visualisation results, demonstrating their advantage in identifying potential symptom phenotypes in sparse binary data.

To assess the model's effectiveness in real-world applications, the study analysed three datasets from the UK: CIS, SGSS, and Pillar2. The symptom sets in SGSS and Pillar2 were consistent, while the community-based CIS dataset covered a broader population but recorded fewer symptoms. Despite differences in sample composition and data collection methods, the core clustering patterns across datasets were highly consistent: fever, cough, headache, and fatigue formed stable clusters, while loss of smell and taste primarily existed as independent modules. Hierarchical clustering and heatmaps further validated the robustness of these patterns. Although

the CIS results had a looser structure, they still revealed stable patterns, indicating the method's strong generalisability.

These three experimental settings fully demonstrate the complementary value of combining simulation and empirical data. Trajectory simulation validated the effectiveness of the method in dynamic environments similar to longitudinal data; binary simulation emphasised interpretability and robustness under sparse conditions; and Jaccard-based analysis established empirical validity across different populations. This hierarchical design achieved a transition from simulated data to real symptom data, thereby enhancing the credibility of the research results. The experimental results indicate that integrating simulation-based and empirical analyses provides a more rigorous framework, which is applicable for revealing symptom co-occurrence patterns.

Methodologically, our study demonstrates the benefits of NMF and UMAP. The additive and non-negative nature of NMF allows results to be straightforward to interpret: each component corresponds to a possible symptom phenotype, which is particularly valuable in a clinical and public health context. By contrast, UMAP effectively balances local and global structures and produces low-dimensional embeddings that are coherent and easily visualised. The combination of both overcomes limitations of PCA when applied to sparse binary matrices and of t-SNE in preserving global structure. Our framework can be applied alongside hierarchical clustering and heatmap visualisation to produce a hierarchical representation of symptom structures. Importantly, using Jaccard similarity enhances robustness, allows for consistent comparison across datasets and provides a common basis for validation.

Despite these advantages, it is important to acknowledge that this study has several limitations. Real-world analyses primarily rely on cross-sectional or aggregated data, which limits the ability to capture the temporal trajectory of symptom evolution. Differences in sampling strategies and population composition across different datasets may also introduce bias in clustering results, thereby limiting the comparability of effects across different cohorts. While parameter sensitivity has been partially addressed in simulation studies, uncertainty remains in real-world applications, which may limit its generalisability. Therefore, future research should expand the current framework in multiple directions. Dynamic methods, such as hidden Markov models or

longitudinal factor models, can more directly capture the temporal evolution of longitudinal symptom data. External validation across different populations and regions is crucial for assessing generalisability. Integrating multimodal data can further enrich the interpretation of symptom co-occurrence mechanisms.

In conclusion, this study systematically evaluated the dimensionality reduction effects of dimensionality reduction and clustering methods on symptom co-occurrence through trajectory simulation, binary symptom simulation, and analysis of real-world Jaccard matrices. The results indicate that the combination of NMF and UMAP is particularly effective in revealing stable and clinically meaningful symptom phenotypes, while Jaccard-based metrics support generalisability across different datasets. This framework is suitable for the methodology of symptom co-occurrence research and provides practical tools for clinical medicine and public health. Despite its limitations, the hierarchical design proposed in this paper lays a solid foundation for future research in longitudinal modelling, multimodal integration, and cross-population validation. With the continuous development of data resources and analytical methods, these methods are expected to play an increasingly important role in deepening our understanding of symptom patterns and disease phenotypes.

References

[1] Aiyegbusi, O.L., Hughes, S.E., Turner, G., Rivera, S.C., McMullan, C., Chandan, J.S., Haroon, S., Price, G., Davies, E.H., Nirantharakumar, K. & Calvert, M.J., 2021. Symptoms, complications and management of long COVID: a review. Journal of the Royal Society of Medicine, 114(9), pp.428–442.

[2] Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W., Ng, L.G., ... & Newell, E.W., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37(1), pp.38–44.

[3] Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P., 2004. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences, 101(12), pp.4164–4169.

[4] Collins, M., Schapire, R.E. & Singer, Y., 2002. Logistic regression, AdaBoost and Bregman distances. Machine Learning, 48(1), pp.253–285.

[5] Floridi, L. & Cowls, J., 2022. A unified framework of five principles for AI in society. In: Machine Learning and the City: Applications in Architecture and Urban Design. Cham: Springer, pp.535–545.

[6] Floridi, L. & Taddeo, M., 2016. What is data ethics?. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), p.20160360.

[7] Fraley, C. & Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), pp.611–631.

[8] Fyles, M., Vihta, K.D., Sudre, C.H., Long, H., Das, R., Jay, C., ... & House, T., 2023. Diversity of symptom phenotypes in SARS-CoV-2 community infections observed in multiple large datasets. Scientific Reports, 13(1), p.21705.

[9] Kriegel, H.P., Kröger, P. & Zimek, A., 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(1), pp.1–58.

[10] Landgraf, A.J. & Lee, Y., 2020. Dimensionality reduction for binary data through the projection of natural parameters. Journal of Multivariate Analysis, 180, p.104668.

[11] Lee, D.D. & Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), pp.788–791.

[12] Lee, D. & Seung, H.S., 2000. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems, 13.

[13] Levandowsky, M. & Winter, D., 1971. Distance between sets. Nature, 234(5323), pp.34–35.

[14] Liu, Y.H., Chen, Y., Wang, Q.H., Wang, L.R., Jiang, L., Yang, Y., ... & Wang, Y.J., 2022. One-year trajectory of cognitive changes in older survivors of COVID-19 in Wuhan, China: a longitudinal cohort study. JAMA Neurology, 79(5), pp.509–517.

[15] Maaten, L.V.D. & Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), pp.2579–2605.

[16] McInnes, L., Healy, J. & Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[17] Menni, C., Valdes, A.M., Freidin, M.B., Sudre, C.H., Nguyen, L.H., Drew, D.A., ... & Spector, T.D., 2020. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature Medicine, 26(7), pp.1037–1040.

[18] Murtagh, F. & Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp.86–97.

[19] Voigt, P. & Von dem Bussche, A., 2017. The EU General Data Protection Regulation (GDPR). A Practical Guide, 1st ed., Cham: Springer International Publishing, 10(3152676), pp.10–5555.

[20] Wolfe, F., Clauw, D.J., Fitzcharles, M.A., Goldenberg, D.L., Häuser, W., Katz, R.L., ... & Walitt, B., 2016. 2016 Revisions to the 2010/2011 fibromyalgia diagnostic criteria. Seminars in Arthritis and Rheumatism, 46(3), pp.319–329.

[21] Xafis, V., Schaefer, G.O., Labude, M.K., Brassington, I., Ballantyne, A., Lim, H.Y., ... & Tai, E.S., 2019. An ethics framework for big data in health and research. Asian Bioethics Review, 11(3), pp.227–254.

.

Appendix

https://github.com/lmc021116/MeichengLiu_ERP