**1.Introduction**

Since the outbreak of the new coronavirus (COVID-19) at the end of 2019, its variable clinical manifestations have posed a significant challenge to the public health response and individualized treatment.The symptoms of COVID-19 are not only diverse, but also show a high degree of heterogeneity across populations, with a wide range of combinatorial patterns, from mild sore throat and malaise to severe respiratory distress and even organ failure (Sudre et al, 2021). As the epidemic has progressed, researchers have begun to realize that it is difficult to accurately identify and respond to potential patient risks based on a single symptom or traditional clinical typing alone.

In recent years, with the accumulation of data from large community-based tracking programs such as the ZOE COVID Symptom Study and CIS (COVID Infection Survey), researchers have been able to analyze the co-occurring structure of symptoms in depth from a data-driven perspective. By using unsupervised learning methods, particularly dimensionality reduction and clustering techniques, scholars have been able to mine the high-dimensional symptom space for potential “symptom phenotypes” that can inform individualized interventions and precise public health policies (Fyles et al., 2023).

The aim of this study is to analyze the underlying structural patterns in the COVID-19 symptom data using a variety of dimensionality reduction and clustering methods, with a particular focus on symptom differences across populations (e.g., age groups) or datasets. The thesis will explore the applicability and interpretability of NMF (non-negative matrix factorization), UMAP (unified manifold approximation and projection), and hierarchical clustering methods in identifying symptom modules and population phenotypes, thus providing methodological support for symptom understanding and disease phenotype modeling.

**2. Theoretical Background**

**2.1 Symptom Co-occurrence and Clinical Phenotypes**

Symptom co-occurrence refers to the phenomenon of an individual experiencing multiple clinical symptoms at the same time during a given disease process. This combinatorial relationship between symptoms may reflect underlying pathological processes, physiological mechanisms, or immune response patterns. In the case of COVID-19, for example, loss of taste and hyposmia are often observed as linked symptoms, the combination of which may be indicative of a specific viral invasion pathway or receptor response (Menni et al., 2020). Phenotyping based on symptom co-occurrence, known as clinical phenotype identification, has the goal of dividing patients into subgroups according to the pattern of symptoms exhibited, which can help personalize treatment and resource allocation (Millar et al., 2022).

**2.2 Application of dimensionality reduction methods in symptom modeling**

Facing symptom data with high dimensionality, high noise, and high correlation between variables, dimensionality reduction techniques have become an important tool for revealing underlying structures. Principal component analysis (PCA), as the earliest linear dimensionality reduction method applied, compresses variable dimensionality while maintaining maximum variance, but its interpretability is weak; nonlinear methods such as t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018) are more suitable for capturing complex, low-dimensional embedded structures, and are particularly suitable for visualizing high-dimensional symptom spaces.UMAP has the advantages of high computational efficiency and relatively low parameter sensitivity while maintaining global structure and local proximity, and thus has been widely adopted in COVID-19 symptom studies (Fyles et al., 2023).

Another important dimensionality reduction method is Non-negative Matrix Factorization (NMF), which extracts the “additive” and “local” feature components by decomposing the original non-negative matrix into two low-dimensional non-negative matrices. localized" feature components by decomposing the original non-negative matrix into two low-dimensional non-negative matrices, thus providing good interpretability (Lee & Seung, 1999). In COVID-19 symptom modeling, NMF is commonly used to identify clinically significant patterns of symptom combinations (Paatero & Tapper, 1994).

**2.3 Clustering methods and phenotype classification**

Cluster analysis, an important branch of unsupervised learning, can automatically identify potential categories or clusters in unlabeled data. In symptom co-occurrence studies, Hierarchical Clustering is commonly used to construct co-occurrence dendrograms of symptoms because it does not require a predetermined number of clusters and has visualization advantages (Kennedy-Shaffer et al., 2022). In addition, methods such as K-means and DBSCAN have also played a role in different studies, especially showing good results in patient clustering and subtype identification tasks. It is worth noting that clustering results are highly dependent on the quality of the input data and the effectiveness of the prior dimensionality reduction process, and thus are often used in conjunction with dimensionality reduction methods to enhance the stability and interpretability of the clustering structure.

**2.4 Analysis of structural changes in multiple populations and time dimensions**

Symptom structure not only varies significantly among individuals, but also among different populations (e.g., the elderly, children, and the vaccinated). Armstrong et al. (2021) pointed out that gender, age, and ethnicity have a systematic influence on the symptom presentation of the COVID-19. Fyles et al. (2023) further applied the “Aligned UMAP” technique to compare and analyze the embedded structures in different datasets, and found that some of the embedded structures are stable across datasets. (2023) further applied the “Aligned UMAP” technique to analyze the embedded structures in different datasets and found that some of the symptom structures were stable across datasets.

In addition, longitudinal data revealed a tendency for symptoms to evolve over the course of the infection. This variability places greater demands on dimensionality reduction and clustering methods to capture temporal dynamics while maintaining interpretability and stability (Greven et al., 2010).

**3. Current status of research**

In recent years, researchers have conducted a number of empirical studies around COVID-19 symptom co-occurrence.Sudre et al. (2021) utilized large-scale user data from the ZOE symptom tracking application to identify six distinct symptom subgroups using a K-means clustering approach, which provided a possible pathway for early clinical risk prediction.Millar et al. (2022), on the other hand, used a hierarchical clustering approach in the UK inpatient ISARIC-4C cohort, used a hierarchical clustering approach to identify seven clinical phenotypes from nearly 60,000 critically ill patients, demonstrating the applicability and clinical relevance of the approach in an inpatient population.

On the other hand, Fyles et al. (2023) used a variety of dimensionality reduction techniques such as UMAP and Logistic PCA on COVID community symptom data (e.g., COVID Symptom Study, CIS, CSS) and found consistent and stable symptom structure across age groups and datasets. Their study not only demonstrated the power of dimensionality reduction methods in revealing the underlying symptom space, but also provided evidence of common structure across datasets. The study also used NMF to further mine interpretable symptom modules, demonstrating the utility of such methods in cluster feature extraction.

In addition, Díaz-Pachón et al. (2022) compared the effectiveness of t-SNE with UMAP in symptom data dimensionality reduction, pointing out that UMAP has a higher performance in preserving the local cluster structure, while Alashkar et al. (2021) introduced UMAP-enhanced clustering to analyze the trajectory of symptom evolution in COVID-19 patients, suggesting that the dimensionality reduction results can assist in the public health response strategy design.

In terms of cross-population comparisons, Armstrong et al. (2021) showed that there were systematic differences in COVID-19 symptom manifestations by gender, age group, and ethnicity, suggesting that symptom co-occurrence studies should focus on hierarchical analyses.Kennedy-Shaffer et al. (2022) clustered symptom combinations across populations through digital symptom surveillance data, further supported the need for structural heterogeneity analysis.