PheMIME: an interactive web app and knowledge base for phenome-wide, multi-institutional multimorbidity analysis

Siwei Zhang, Nick Strayer, PhD, Tess Vessels, BS, Karmel Choi, PhD, Geoffrey W. Wang, BS, Yajing Li, MS, Cosmin A. Bejan, PhD, Ryan S. Hsi, MD, Alexander G. Bick, MD, PhD, Digna R. Velez Edwards, PhD, Michael R. Savona, MD, Elizabeth J. Phillips, MD, Jill M. Pulley, MBA, Wesley H. Self, MD, MPH, Wilkins Consuelo Hopkins, MD, Dan M. Roden, MD, Jordan W. Smoller, MD, Douglas M. Ruderfer, PhD, Yaomin Xu, PhD

1Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, 2Posit PBC, Boston, MA 02210, United States, 3Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 4Psychiatric & Neuro Developmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, United States, 5Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114, United States, 6Department of Statistics, North Carolina State University, Raleigh, NC 27695, United States, 7Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, 8Department of Urology, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 9Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 10Division of Hematology and Oncology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 11Center for Drug Safety and Immunology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 12Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, WA 6150, Australia, 13Vanderbilt Institute for Clinical and Translational Science, Vanderbilt University Medical Center, Nashville, TN 37203, United States, 14Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN 37232, United States, 15Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA 02142, United States, 16Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN 37212, United States

*Corresponding author: Yaomin Xu, PhD, Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 1100, Nashville, TN 37203, United States (yaomin.xu@vumc.org)

Abstract

Objectives: To address the need for interactive visualization tools and databases in characterizing multimorbidity patterns across different populations, we developed the Phenome-wide Multi-Institutional Multimorbidity Explorer (PheMIME). This tool leverages three large-scale EHR systems to facilitate efficient analysis and visualization of disease multimorbidity, aiming to reveal both robust and novel disease associations that are consistent across different systems and to provide insight for enhancing personalized healthcare strategies.

Materials and Methods: PheMIME integrates summary statistics from phenome-wide analyses of disease multimorbidities, utilizing data from Vanderbilt University Medical Center, Mass General Brigham, and the UK Biobank. It offers interactive and multifaceted visualizations for exploring multimorbidity. Incorporating an enhanced version of associationSubgraphs, PheMIME also enables dynamic analysis and inference of disease clusters, promoting the discovery of complex multimorbidity patterns. A case study on schizophrenia demonstrates its capability for interpreting the complex and noisy real-world patient data in electronic health records.

Results: The PheMIME is accessible at https://prod.tbilab.org/PheMIME/. A comprehensive tutorial and multiple case studies for demonstration are available at https://prod.tbilab.org/PheMIME_supplementary_materials/. The source code can be downloaded from https://github.com/tbilab/PheMIME.

Discussion: PheMIME represents a significant advancement in medical informatics, offering an efficient solution for accessing, analyzing, and interpreting the complex and noisy real-world patient data in electronic health records.

Conclusion: PheMIME provides an extensive multimorbidity knowledge base that consolidates data from three EHR systems, and it is a novel interactive tool designed to analyze and visualize multimorbidities across multiple EHR datasets. It stands out as the first of its kind to offer extensive multimorbidity/knowledge integration with substantial support for efficient online analysis and interactive visualization.

Key words: PheWAS; multimorbidity; interactive network analysis; UK Biobank; interoperability and reproducibility; electronic health records (EHRs).

Introduction

The global rise of multimorbidity, the simultaneous occurrence of several diseases within an individual, poses a significant challenge to public health and healthcare providers. In-depth understanding of intricate multimorbidity can uncover the complex interactions between diseases that affect their course, severity, and the patient’s treatment response. Moreover, it may reveal shared molecular mechanisms among various diseases, offering opportunities for innovative prevention strategies, targeted...
interactions, and personalized treatments for patients with multimorbidities.\textsuperscript{12–15}

Large-scale electronic health record (EHR) systems, linked to comprehensive health and clinical data, have significantly enhanced the statistical power to examine robust multimorbidity patterns that mirror real-world scenarios.\textsuperscript{7,11,16–19} Network analysis has been a powerful approach to decipher complex multimorbidity patterns.\textsuperscript{20–22} Our recent work has demonstrated the interoperability of EHR-based multimorbidities and multimorbidity networks when compared across distinct EHR systems using standardized diagnostic codes (eg, ICD9 or ICD10).\textsuperscript{23,24}

Integrating multiple EHR systems can significantly improve the generalizability and robustness of multimorbidity characterization by identifying commonalities and differences across various populations. However, current standards for quantifying, representing, and analyzing multimorbidity patterns, especially for phenome-wide analyses leveraging large-scale EHRs, are still lacking.\textsuperscript{1,23}

In this study, we introduce the Phenome-wide Multi- Institutional Multimorbidity Explorer (PheMIME), an interactive web application developed using the R programming language and the Shiny library.\textsuperscript{26} PheMIME enables researchers to interactively explore, compare, and discover multimorbidity patterns using a phenome-wide multimorbidity knowledge base compiled from three major EHR databases: Vanderbilt University Medical Center (VUMC), Mass General Brigham Hospital (MGB), and UK Biobank (UKB).\textsuperscript{24} With the selection of a specific disease phenotype of interest, users can directly compare multimorbidities between different institutions, conduct a network analysis based on individual or multi-institution combined multimorbidity network, and visualize dynamic clustering of significant multimorbidities by applying filters.

PheMIME provides a rich, explorable view of the multimorbidity patterns and relationships among diseases across multiple institutions. It can facilitate the validation of novel disease associations, the discovery of robust disease multimorbidity patterns, and potentially reveal new insights for future investigation. We demonstrate PheMIME’s analytical capabilities and customization options through a schizophrenia case study. Additional detailed case studies available online further showcase diverse multimorbidity-based discovery schemes.

**Methods**

**Data integration and multimorbidity knowledge base**

For our analysis, we accessed three EHR databases that included individual-level data for 250,000 random patients each from the VUMC and MGB EHR systems, as well as data from 431,105 subjects in the UKB. For UKB, in-hospital records field-41270, which records the distinct diagnosis codes (ICD 10) a participant has had recorded across all the hospital inpatient records in either the primary or secondary position, was used for the analysis. The ICD 10 codes were then mapped to phecodes\textsuperscript{27,28} where a phecode was denoted as an occurrence for an individual if it appeared at least once in a UKB subject’s collapsed record across all hospital inpatient records, or if occurred two or more times in a patient’s record for either VUMC or MGB. Specifically, a total of 1815 phecodes were identified across the dataset. Broken down by resource, VUMC, MGB, and UKB reported 1815, 1805 and 1484 phecodes, respectively. The most significant factor contributing to the variation in phecode counts is the difference in population characteristics between institutions. VUMC and MGB primarily serve as hospitals, where subjects are predominantly patients seeking medical care. In contrast, the UKB represents a general population cohort, comprising individuals from the community who may or may not have a history of medical conditions.\textsuperscript{25} Despite these variations, our comprehensive analysis revealed in total 1 485 133 pairs of phecodes that demonstrated overlap and relationships, highlighting the connections between different phecodes within our datasets. Logistic regressions adjusting for patient age at last recorded visit, sex, race, and the number of unique phecodes present in patients’ records were run for each pair of two phecodes with two conditions of either phecode A or phecode B regarded as the outcome and the other one treated as the independent variable.\textsuperscript{23,24} The averaged test statistic from the two regression analyses is then used as an estimate for the multimorbidity strength between the phecode pair A and B. Multimorbidity strengths of all pairs are subsequently calculated and used to construct a phenome-wide multimorbidity network, which represents an undirected weighted network with disease as nodes and disease-disease connections as edges, weighted by the multimorbidity strengths. In addition, Pearson correlation of the common multimorbidity patterns between each pair of phecodes is used as a similarity score, which was also used to generate another undirected weighted network using the similarity scores as weights with disease as nodes and connections as edges.\textsuperscript{24} We call this a multimorbidity similarity network. We observed that the similarities in multimorbidity patterns among diseases significantly correlate with their genetic correlations,\textsuperscript{24} suggesting that multimorbidity similarity can serve as a valuable indicator of potential shared etiology. We finally consolidate all the summarized data from three institutions into a database. This serves as a knowledge base to fuel PheMIME for phenome-wide multimorbidity analysis.

**Design scheme**

PheMIME is an interactive web-based tool designed to enable clinical researchers to explore complex multimorbidity patterns across large-scale EHR data, translating robust findings into actionable insights. Strayer et al demonstrated the interoperability of network analysis methods for phenome-wide multimorbidity data.\textsuperscript{24} Building upon this, PheMIME addresses the challenges of analyzing extensive multimorbidity datasets while enabling focused exploration relevant to specific clinical areas or context. It harnesses the advantages of a phenome-wide approach, using phecodes to map to real-world clinical vocabularies, which enhances interpretability. Users can select phenotypes of interest and investigate complex patterns in a comprehensive and customizable manner. Its dual approach, combining broad-scale visualization with deeper dives into specific multimorbidity patterns, facilitates the efficient extraction of meaningful insights. Additionally, PheMIME supports researchers to validate findings across different institutions, illustrating their reproducibility.

PheMIME incorporates five primary modules: (1) “Disease Selection” module facilitates an interactive phecode table where users can readily search, filter, and select a disease phecode of interest. (2) “Multimorbidity Consistency Inspection” module enables users to assess the overall
consistency of multimorbidity strength measurements from the knowledge base and compare them across multiple institutions. Additionally, this module incorporates features that underscore significant multimorbidity strengths linked to the selected phecode of interest, aids in assessing their distribution amidst all multimorbidity measurements, and enables comparison across institutions. (3) “Multimorbidity Network Visualization” module presents interactive visual representations of the multimorbidity networks constructed based on the multimorbidity strength measurements and allows users to examine and compare these networks. This module integrates associationSubgraphs as an advanced tool for dynamic network clustering and visualization (detailed in the “Dynamic Network Clustering of Multimorbidities With AssociationSubgraphs” section). This module permits exploration of the network’s subgraph structures and dynamic clustering for any multimorbidity network from a single institution or an amalgamation of multiple institutions. Moreover, this module enables users to apply filters and emphasize any significant multimorbidities and investigate their interconnections and enriched subgraphs. (4) “Reproducible Multimorbidities Exploration” module provides an interactive environment for examining a customizable subset of significant and/or reproducible multimorbidities across the institutions based on the multimorbidity strength measurements hosted in the knowledge base. This interface allows users to visualize the interconnections among chosen phecodes and the enriched multimorbidity subgraphs within the combined multimorbidity networks. Furthermore, this module accommodates pairwise comparisons between all institutions. (5) “Multimorbidity Similarities Exploration” module, much like the preceding one, uses multimorbidity similarity measurements as the strength measurement. It permits visualization of interconnected phecodes and the multimorbidity subgraphs enriched in the combined multimorbidity similarity networks and enables pairwise comparisons between all institutions.

PheMIME is built using the R programming language, the JavaScript D3.js library, and the Shiny library. To manage extensive computations, we employ parallel processing and cloud computing resources (AWS EC2 instance, 64 cores, 500 GB memory). Data storage is handled through AWS S3. The platform dynamically recalculates the network based on user input, allowing real-time customization of network structure and enabling flexible analyses tailored to specific multimorbidity discovery needs.

Dynamic network clustering of multimorbidities with associationSubgraphs
In PheMIME, we integrate a dynamic network clustering technique called the associationSubgraphs. This is an interactive clustering approach that identifies subgraphs, or disease clusters, in multimorbidity networks by leveraging the principles of threshold graph construction. The algorithm initiates by sorting the edges of the network, which represent disease-disease connections, in descending order of their comorbidity strength. The strongest link forms the core of the initial cluster, with subsequent edges added iteratively. As each edge is incorporated, its nodes are evaluated: if one node is already part of a cluster, the other node joins the same cluster; if both nodes are unconnected, they start a new cluster; if each node belongs to a different cluster, those clusters merge. The algorithm meticulously records the state of clusters after each addition, enabling the complete tracking of how clusters evolve over time.

To establish the optimal point for halting the clustering process, we apply the “largest-smallest” rule, which identifies when well-defined clusters begin transitioning into a large, connected component, indicating the onset of structural randomness. For further details on this methodological approach, see Strayer et al.

Our implementation of the associationSubgraphs algorithm demonstrates its effectiveness in dynamically clustering diseases based on varying comorbidity strengths. This method offers nuanced insights into the hierarchical structure of disease clusters and their patterns of amalgamation at different comorbidity thresholds. Importantly, this process identifies expansive clusters even in scenarios where diseases do not show densely connected or uniformly strong pairwise comorbidities. Such capabilities distinguish this technique from traditional community detection methods, which may falter under these sparse conditions.

Results

Analysis outline using schizophrenia as an example
This section outlines the workflow of PheMIME to explore phenome-wide multimorbidity patterns among various analysis modules provided. Starting with phecode 295.10 (schizophrenia), PheMIME generates an interactive Manhattan plot, enabling users to select phecodes based on their pairwise multimorbidity strength, also known as comorbidity strength (Figure 1A). A scatter plot for comparing between two systems is also generated, helping users to select robust multimorbidities that exhibit reproducible patterns across systems (Figure 1A). The interactive features of the Manhattan and Scatter plots enable the selection and highlighting of a consistent set of the same phecodes, based on the magnitude (Manhattan plot) and consistency (scatter plot) of disease multimorbidities. For instance, the Manhattan plot might reveal phecodes highly co-occurring with schizophrenia, whereas the scatter plot could highlight phecodes that are either highly consistent across institutions or remain exclusive to individual institutions. The data table in Figure 1B shows comorbid phecodes of schizophrenia, its description, disease categories and corresponding multimorbidity strengths among three institutions. This table is interactive with the Manhattan and Scatter plots, allowing users to add or remove phecodes by clicking on the rows in the table. If a disease multimorbidity exhibits both a large magnitude and high consistency across different systems, it strongly indicates a robust disease multimorbidity across the systems.

For dynamic network analysis, the associationSubgraphs method has been enhanced to provide an interactive visualization to rapidly explore subgraph structures of schizophrenia multimorbidities in the selected phecodes that are both significant and reproducible across institutions (filled circles in Figure 1A and B). As shown in Figure 1C, network nodes are annotated into two groups, with the selected phecodes color-filled based on disease categories and the other selected phecodes (nodes) color-filled in grey. As expected, a major subgraph enriching the selected schizophrenia multimorbidity phecodes is mental disorders, with other subgraphs showing enrichment in infectious diseases and neoplasms.

PheMIME also provides an expedient and accessible method for comparative analyses of multimorbidity patterns.
across different populations. Using schizophrenia again as an example, we discovered a strong comorbidity with viral hepatitis B and C (as detailed in the online materials), a finding in line with prior research.\textsuperscript{30,31} Yet, a detailed examination into multimorbidity patterns revealed more consistency of comorbidity within the patient cohorts of VUMC and MGB, with the comorbidity intensities being notably stronger in these cohorts as compared to the general UKB cohort, especially in the case of viral hepatitis B. This distinction could underscore heterogeneity due to population differences—
consistent with the observation of higher prevalence of blood-borne viral infections for patients with serious mental illness compared to the general population. However, the exact mechanisms behind viral hepatitis and schizophrenia comorbidity remain elusive, as observed disparities among different cohorts could motivate additional investigation regarding the mechanisms and impact of population differences on disease multimorbidity patterns.

Exploring context-specific phenome-wide disease clusters
We implemented the associationSubgraphs algorithm in PheMIME to identify disease clusters in multimorbidity networks. This method enables systematic discovery of disease clusters by allowing users to dynamically adjust the threshold for each unique comorbidity strength value, providing a complete view of hierarchical clustering process across all levels of comorbidity strengths from low to high. The effectiveness of associationSubgraphs has been demonstrated by its ability to identify prominent disease condition clusters consistent with previously reported findings. For further details, refer to the online Supplementary material—“Case Study: Identification of Robust Disease Condition Clusters”, and Strayer et al.

Here, we illustrate how PheMIME facilitates the exploration of robust disease clusters provided a specific disease phenotype of interest, using statistical significance and multimorbidity correlation as criteria. Using schizophrenia as an example, users can uncover the following insights:

1) Identify multimorbidities associated with schizophrenia across different systems and explore disease clusters at specified statistical significance levels through the Multimorbidity Network Visualization module.
2) Identify consistent disease phenotypes based on comorbidity strength as shown in the Manhattan and Scatter plot through the Reproducible Multimorbidity Exploration module and visualize the disease subgraphs using associationSubgraphs.
3) Similarly, identify consistent disease phenotypes based on multimorbidity similarities in the Manhattan and Scatter plot through the Multimorbidity Similarities Exploration module and visualize the disease subgraphs using associationSubgraphs.
4) Supplement these findings with arbitrary disease phenotypes of interest by selecting them from the customizable table through the Reproducible Multimorbidities Exploration or Multimorbidity Similarities Exploration modules. Incorporate them into the final selection by clicking the Update Manhattan/Scatter button.

associationSubgraphs analysis enables the identification of clinically relevant disease clusters within specific disease contexts. Prior research suggests that individuals with schizophrenia may have a slightly elevated risk for certain cancers, such as colon and breast cancer. Our findings confirm the enrichment of cancer-related subgraphs. Figure 2 displays the consistent association of these multimorbidities with schizophrenia across multiple systems, with significant comorbidity defined at a P value cutoff of .05. Network nodes, color-coded by disease categories, highlight phenotypes significantly associated with schizophrenia, including colon, breast, and prostate cancer, known to correlate with schizophrenia. Additionally, we observed an unusual pattern with skin cancer within this subgraph, rarely explored in previous studies, potentially offering novel insights.

Discussion
The development of PheMIME significantly enhances research in exploring and comparing multimorbidity patterns across multiple institutions. This tool meets the critical need for comprehensive analysis of multimorbidity patterns by integrating data from various institutions, thereby transforming the approach of researchers and clinicians to the study of multimorbidity diseases. A key strength of PheMIME lies in its ability to integrate and visualize large-scale EHR data from diverse populations, which is crucial for developing effective prevention strategies and targeted interventions. Moreover, it offers efficient, user-customizable online visualization, enabling researchers to perform dynamic analysis and inference of disease clusters.

In this study, we utilize phecodes to harness the benefits of a phenome-wide approach, aligning with our goal to enable straightforward interpretation by clinicians engaged in data-driven discoveries. Phecodes map ICD codes into clinically meaningful categories, facilitating the exploration of
disease relationships across the entire phenome. While specialized solutions like OHDSI’s Phenotype Library and PheKB offer greater precision, our phenocode-based approach prioritizes a user-friendly interface and direct links to real-world clinical vocabularies. This design choice improves PheMIME’s usability and potential for EHR data-driven clinical translational research and continuous learning in healthcare, especially for researchers who may find more complex solutions less accessible.

However, our study has limitations. While PheMIME encompasses phenome-wide analysis, exploring disease-disease networks interconnected through common genes, proteins, or metabolites is also vital. This exploration could uncover shared biomolecular mechanisms, potentially identifying targets for drug repurposing and contributing to precision medicine. Additionally, the current dataset, though extensive, is limited to three major institutions, potentially impacting the generalizability of our findings. Future enhancements of PheMIME could involve expanding the database to encompass more diverse populations and healthcare settings, thus augmenting the robustness of the analysis.

Conclusion
We have introduced PheMIME, an interactive visualization tool specifically designed for analyzing multimorbidities across multiple EHR datasets, which simultaneously presents an extensive multimorbidity knowledge base consolidating data from three major EHR systems. Our intention is for users to utilize PheMIME to detect and extract meaningful disease multimorbidities, and to compare and validate them across a variety of institutions. To our understanding, PheMIME is the first knowledge base of its kind, integrating and comparing data from multiple extensive EHR systems while providing substantial support for efficient online analysis and interactive visualization to facilitate the discovery of complex multimorbidity patterns.

Acknowledgments
We are deeply grateful for the insightful discussions with Kedir N. Turi, Sharon E. Phillips, Xiaopeng Sun, Lydia Yao, Eric Chen, Brian Sharber, and Matt Krantz, which significantly enriched the development and direction of this work. Their input and expertise were invaluable and greatly appreciated.

Author contributions
Yaomin Xu conceived the idea and designed the study, Siwei Zhang and Nick Strayer primarily conducted the major statistical analysis and played a key role in the design and development of the web application. Tess Vessels provided assistance in data analysis and contributed significantly to the interpretation of the genetic data analysis. Siwei Zhang prepared the initial draft of the manuscript, and Yaomin Xu was crucial in its revision. Yaomin Xu and Douglas M. Rudfer offered significant guidance in both the design of the study and the data collection process. All authors reviewed, provided valuable feedback, and gave their final approval for the published version of the manuscript.

Funding
N.S. and Y.X. are supported by the Vanderbilt University Department of Biostatistics Development Award; Y.X., C.B., and R.H. are supported by R21DK127075; Y.X., D.E., E.P., and D.R. are supported by P50GM115305; J.W.S. is supported in part by R01 MH118233. The Vanderbilt University Medical Center dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center’s SD/BioVU, which is supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vanderbilt.edu/pub/biovu/. This research has been conducted using the UK Biobank Resource under Application Number 43397.

Conflicts of interest
None declared.

Data availability
Individual-level data are not shared due to ethical and privacy constraints. However, all source code used in the analysis is publicly available on our GitHub repository at https://github.com/tbilab/PheMIME. Additionally, summary data from the multimorbidity analyses can be accessed through the PheMIME App, available at https://prod.tbilab.org/PheMIME/.

Author declarations
All relevant ethical guidelines have been followed and any necessary IRB and/or ethics committee approvals have been obtained. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived. I have followed all appropriate research reporting guidelines and uploaded the relevant Equator, ICMJE, or other checklist(s) as supplementary files, if applicable.

References


