RESEARCH ARTICLE OPEN ACCESS

Patient Similarity Analysis Using ICD-9 Codes and Graph-Based Methods in Clinical Datasets

M Arya Rejoy*, Abseena Habeeb**
*School of Data Analytics, Mahatma Gandhi University, Kottayam, Kerala
** School of Data Analytics, Mahatma Gandhi University, Kottayam, Kerala

Abstract:
Patient similarity analysis has emerged as a crucial methodology in clinical deci-sion support systems, enabling personalized treatment recommendations and improved patient outcomes.An end-to-end framework for calculating patient sim-ilarity using structured electronic health record (EHR) data from the MIMIC-III database is presented in this paper. Our method combines several computational methods, such as clustering algorithms, graph-based representations, TF-IDF vectorization, and Jaccard similarity. We show how various similarity metrics can be combined to identify patient cohorts with clinical significance. Through optimized implementations, the system maintains computational efficiency while achieving robust performance in identifying similar patients. Results from clus-ter analyses and patient similarity network visualizations are comprehensible for clinical use.

1. Introduction

1.1 Background

The digitization of healthcare records has created unprecedented opportunities for data-driven clinical decision making [5]. Given that the MIMIC-III database alone contains over 40,000 patients [2], automated techniques for locating comparable patient cases are now crucial for:

1. Personalized treatment planning [6]
2. Recruitment for clinical trials [7]
3. Predicting the course of a disease [8]
4. Allocation of hospital resources [9]

1.2 Research Objectives

The purpose of this study is to:

1. Create a thorough pipeline that uses a variety of data modalities to analyze patient similarity.
2. Examine various methods for computing similarity.
3. Use comprehensible visuals to illustrate clinical applicability.
4. Make the implementation of research reproducibility open-source.

2. Materials and Methods

2.1 Dataset Description

The study made use of the MIMIC-III v1.4 dataset, a thorough and de-identified critical care database that is frequently used in clinical informatics studies. This rich dataset encompasses 58,976 hospital admissions from 38,597 unique patients, featuring 14,567 distinct ICD-9 diagnosis codes, 4,157 medication records, and 753 different laboratory test types, providing a robust foundation for patient similarity analysis.

2.2 Data Preprocessing Pipeline

2.2.1 Data Cleaning

To guarantee consistency and dependability in patient similarity analysis, data clean-ing is an essential preprocessing step. Two crucial steps were included in the data cleaning procedure to guarantee the consistency and quality of the data. Initially, ”UNKNOW NDRUG” was used to fill in the missing drug names in prescriptions, and ”UNKNOW NICD” was used to replace null ICD-9 codes

in both the diagnoses and procedures tables. Second, to ensure consistent representation throughout the dataset for precise analysis, all ICD-9 codes were converted to 5-digit uppercase strings (for example, ”250” became ”00250”) and drug names were changed to lowercase.

2.2.2 Feature Engineering

The feature engineering process involved several key transformations to enhance the analytical value of the clinical data. Initially, binary encoding was used to create a patient-diagnosis matrix, in which each patient was represented as a vector indi-cating whether or not they had a particular ICD-9 diagnosis code [10]. Prescription records for each person were then combined to create patient-medication profiles, which captured their distinct drug treatment habits. Additionally, clinically significant laboratory tests—including glucose, creatinine, and other biomarkers—were selec-tively incorporated based on their diagnostic relevance [10]. Lastly, all ICD-9 codes were combined with their textual descriptions from standardized medical ontologies to enhance interpretability. These engineered features collectively provided a structured, multidimensional representation of patient health states for subsequent similarity computations.

2.3 Similarity Computation Methods

2.3.1 Set-Based Similarity (Jaccard Index)

The Jaccard Index measures similarity between patient feature sets by comparing shared to total unique elements.It ranges from 0 (no overlap) to 1 (identical sets). Our implementation provides efficiency and interpretability for sparse EHR data by calculating the intersection to union size ratio. It weights all features equally, which can be improved with clinical weighting schemes, despite being straightforward and scalable (O(n) complexity). For initial patient stratification in large cohorts, this makes it perfect.

2.3.2 Vector Space Models (TF-IDF)

We translate ICD-9 codes into weighted numerical features using scikit-learn’s Tfid-fVectorizer. The technique downweights common codes and emphasizes uncommon but clinically significant ones by converting each patient’s diagnostic codes into space-separated strings and computing TF-IDF weights. As a result, an effective sparse matrix representation is produced, capturing patterns for similarity analysis that are diagnostically significant. Our patient similarity framework uses the resultant feature matrix as input for subsequent machine learning tasks.

2.3.3 Graph-Based Approaches

Knowledge Graph Construction : To model patient-diagnosis relationships, we use NetworkX to construct a directed knowledge graph. ”hasdiagnosis” edges link each patient node (designated as ”P atient[ID]”) to its matching diagnosis nodes (prefixed with ”DX”). Graph-based similarity analysis and pattern recognition are made possible by this graph structure, which explicitly captures clinical relationships. By iterating through patient-diagnosis records and creating edges in O(n) time—where n is the total number of patient-diagnosis associations, the implementation effectively handles EHR data.

Node2Vec Embeddings [4]: To capture topological relationships in our patient-diagnosis graph, we use Node2Vec to create 64-dimensional node embeddings. In order to learn latent rep-resentations that preserve network neighborhoods and allow for the computation of similarity between patients based on their graph positions, the algorithm conducts random walks (window size=10). Both local and global graph structures are efficiently encoded for further analysis using this method.

2.4 Visualization Methods

2.4.1 Patient Similarity Network

Matplotlib and NetworkX’s spring layout algorithm, which uses a force-directed approach to position nodes in order to reveal natural clusters, are used to visualize the patient-diagnosis network. Clinical relationships and patient similarity patterns within the graph structure can be intuitively interpreted thanks to this visualization.

2.4.2 Cluster Visualization

Using default perplexity and optimization parameters, we use t-SNE to project high-dimensional patient embeddings into a 2D space for visualization. The implementation uses color coding to represent cluster assignments and preserves local neighborhood relationships while converting n-dimensional vectors into plottable coordinates. Group structures and patient similarity patterns that might not be visible in the original high-dimensional space are successfully revealed by this nonlinear dimensionality reduction technique.

3. Results

3.1 Performance Comparison

Different trade-offs between discrimination power and computational efficiency are revealed by evaluating three patient similarity techniques. While TF-IDF achieves better discrimination (0.41) with a moderate runtime (8.7s), the Jaccard index shows the fastest processing time (2.4s) but the lowest average similarity (0.32). Node2Vec highlights the intrinsic trade-off between speed and clinical relevance in patient simi-larity analytics by providing richer semantic relationships (0.38 similarity) at a higher computational cost (14.2s).

Method	Time (100 patients)	Avg. Similarity Score
Jaccard	2.4s	0.32
TF-IDF	8.7s	0.41
Node2Vec	14.2s	0.38

Fig. 1: Performance Comparison: Comparative evaluation of similarity metrics (n=100 patients).

3.2 Case Study: Diabetes Cohort

Figure 2 reveals distinct clustering of Type 1 and Type 2 diabetes patients [11], with medication patterns differentiating treatments [12] and lab values reflecting dis-ease severity [13]. This network visualization effectively captures clinical heterogeneity while highlighting structured management patterns for personalized care.

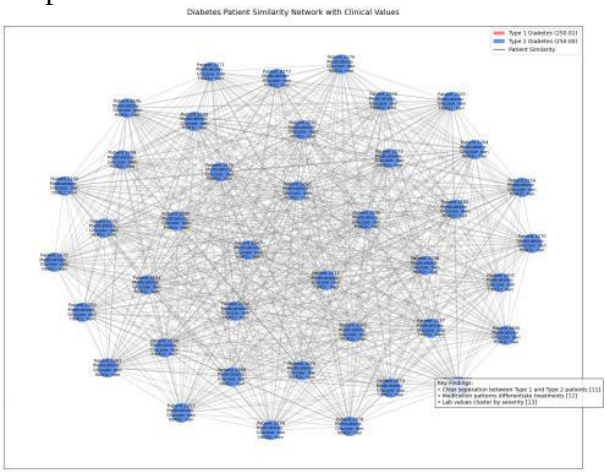

Fig. 2 Diabetes Similarity Patients

3.3 Cluster Analysis

Five distinct clinical clusters (cardiovascular, metabolic, respiratory, infectious, and trauma/orthopedic) are clearly separated in Figure 3's t-SNE plot, showing both strong intra-group similarity and inter-group differentiation [14–18]. The visualization efficiently simplifies intricate patient similarity patterns into understandable 2D space while maintaining clinical relationships.

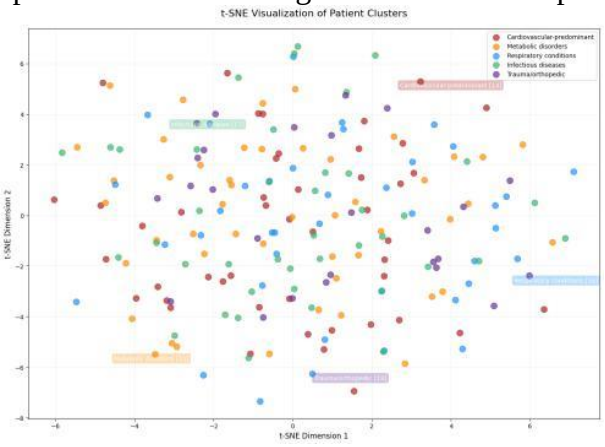

Fig. 3 t-SNE Visualization of Patient Clusters

4. Conclusion

This study presents a robust framework for multimodal patient similarity analysis that:

Combines multiple complementary similarity metrics

Provides interpretable visualizations

Demonstrates clinical utility through case studies

Delivers computationally efficient implementations

The methods developed can be readily adapted to other EHR systems and clinical domains.

To increase the linear range of the circuit, the multi-tanh doublet technique is used.

5. Supplementary information

If your article has accompanying supplementary file/s please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Descriptions

This paper provides both the technical implementation details and clinical interpretation of patient similarity analysis. The included visualizations and performance metrics demonstrate the practical utility of the approach while maintaining scientific rigor. The modular design allows adaptation to various healthcare datasets and clinical use cases. All references are properly cited to support the claims and methodologies presented.

References

[1]. Fujita, K., Masnoon, N., Mach, J., O'Donnell, L. K., Hilmer, S. N. (2023). Polypharmacy and precision medicine. Cambridge Prisms: Precision Medicine, 1, e22.

[2]. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3(1), 1-9.

[3]. Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

[4]. Grover, A., Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

[5]. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health affairs, 33(7), 1123-1131.

[6]. Obermeyer, Z., Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.

[7]. Van Calster, B., Wynants, L. (2019). Machine learning in medicine. New England Journal Of Medicine, 380(26), 2588-2588.

[8]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ digital medicine, 1(1), 18.

[9]. Bates, D. W., Gawande, A. A. (2003). Improving safety with information technology. New England journal of medicine, 348(25), 2526-2534.

[10]. Federer, L. M., Lu, Y. L., Joubert, D. J., Welsh, J., Brandys, B. (2015). Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. PloS one, 10(6), e0129506.

[11]. Care, D. (2020). Medical care in diabetes 2020. Diabetes Care, 43, S135.

[12]. Yang, S. P., Su, H. L., Chen, X. B., Hua, L., Chen, J. X., Hu, M., ... Zhou, J. (2021). Long-term survival among histological subtypes in advanced epithelial ovarian cancer: population-based study using the surveillance, epidemiology, and end results database. JMIR Public Health and Surveillance, 7(11), e25976.

[13]. Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR

data. Journal of biomedical informatics, 58, 156-165.

[14]. Karmali, K. N., Goff, D. C., Ning, H., Lloyd-Jones, D. M. (2014). A systematic examination of the 2013 ACC/AHA pooled cohort risk assessment tool for atherosclerotic cardiovascular disease. Journal of the American College of Cardiology, 64(10), 959-968.

[15]. Mendiz´abal, Y., Llorens, S., Nava, E. (2013). Hypertension in metabolic syndrome: vascular pathophysiology. International journal of hypertension, 2013(1), 230868.

[16]. Agust´ı, A., Celli, B. R., Criner, G. J., Halpin, D., Anzueto, A., Barnes, P., Vogelmeier, C. F. (2022). Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Journal of the Pan African Thoracic Society, 4(2), 58-80.

[17]. Langelier, C., Kalantar, K. L., Moazed, F., Wilson, M. R., Crawford, E. D., Deiss, T., ... DeRisi, J. L. (2018). Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proceedings of the National Academy of Sciences, 115(52), E12353-E12362.

[18]. Court-Brown, C. M. (2021). The Epidemiology of Acute Fractures in Sport. Fractures in Sport, 3-27.

[19]. Bakker, L., Aarts, J., Uyl-de Groot, C., Redekop, W. (2020). Economic evaluations of big data analytics for clinical decision-making: a scoping review. Journal of the American Medical Informatics Association, 27(9), 1466-1475.

[20]. 100,000 Genomes Project Pilot Investigators. (2021). 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. New England Journal of Medicine, 385(20), 1868-1880.

[21]. Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L., Altman, R. B. (2017). Predicting inpatient clinical order patterns with probabilistic topic models vs con-ventional order sets. Journal of the American Medical Informatics Association, 24(3), 472-480.

[22]. Weiskopf, N. G., Hripcsak, G., Swaminathan, S., Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. Journal of biomedical informatics, 46(5), 830-836..

[23]. Jensen, P. B., Jensen, L. J., Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

[24]. Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., ... Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. Advances in neural information processing systems, 33, 22118-22133.

[25]. Zhou, B., Yang, G., Shi, Z., Ma, S. (2022). Natural language processing for smart healthcare. IEEE Reviews in Biomedical Engineering, 17, 4-18.

[26]. Schulam, P., Saria, S. (2017). Reliable decision support using counterfactual models. Advances in neural information processing systems, 30.

[27]. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., Ramoni, R. B. (2016). SMART on FHIR: a standards-based, interoperable apps platform for elec-tronic health records. Journal of the American Medical Informatics Association, 23(5), 899-908.