Journal of the American Medical Informatics Association, 2025, 1–9
https://doi.org/10.1093/jamia/ocae314
Research and Applications

Research and Applications

Utility of word embeddings from large language models in medical diagnosis

Shahram Yazdani, MD*,[1], Ronald Claude Henry, PhD[2], Avery Byrne, BA, BS[3], Isaac Claude Henry, BS, MS[4]

[1]Department of Pediatrics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, United States, [2]Department of Civil Engineering, University of Southern California, Los Angeles, CA 90089, United States, [3]San Francisco, CA 94129, United States, [4]Kennewick, WA 99338, United States

*Corresponding author: Shahram Yazdani, MD, Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, 10833 Le Conte Ave, Los Angeles, CA 90095-3075, United States (syazdani@mednet.ucla.edu)

Abstract

Objective: This study evaluates the utility of word embeddings, generated by large language models (LLMs), for medical diagnosis by comparing the semantic proximity of symptoms to their eponymic disease embedding ("eponymic condition") and the mean of all symptom embeddings associated with a disease ("ensemble mean").

Materials and Methods: Symptom data for 5 diagnostically challenging pediatric diseases—CHARGE syndrome, Cowden disease, POEMS syndrome, Rheumatic fever, and Tuberous sclerosis—were collected from PubMed. Using the Ada-002 embedding model, disease names and symptoms were translated into vector representations in a high-dimensional space. Euclidean and Chebyshev distance metrics were used to classify symptoms based on their proximity to both the eponymic condition and the ensemble mean of the condition's symptoms.

Results: The ensemble mean approach showed significantly higher classification accuracy, correctly classifying between 80% (Cowden disease) to 100% (Tuberous sclerosis) of the sample disease symptoms using the Euclidean distance metric. In contrast, the eponymic condition approach using Euclidian distance metric and Chebyshev distances, in general, showed poor symptom classification performance, with erratic results (0%-100% accuracy), largely ranging between 0% and 3% accuracy.

Discussion: The ensemble mean captures a disease's collective symptom profile, providing a more nuanced representation than the disease name alone. However, some misclassifications were due to superficial semantic similarities, highlighting the need for LLM models trained on medical corpora.

Conclusion: The ensemble mean of symptom embeddings improves classification accuracy over the eponymic condition approach. Future efforts should focus on medical-specific training of LLMs to enhance their diagnostic accuracy and clinical utility.

Key words: natural language processing; machine learning; pediatric diagnosis; word embeddings; clinical decision support.

Background and significance

Meticulous attention to medical terminology in both clinical practice and medical education, along with the development of standardized lexicons for describing anatomy, pathology, and symptoms, underscores the deep-rooted role of semantics in medicine. This precise use of language is a key reason for the growing interest in applying advanced natural language processing (NLP) techniques, particularly word embeddings, to clinical practice.[1] Word embeddings, a deep learning technique, encapsulate words as vectors within a high-dimensional space, creating a machine-interpretable semantic representation of words. This method allows for the quantification of contextual proximity between terms, offering novel opportunities for enhancing clinical diagnosis and practice.[2–4]

Since word embeddings are generated by deep learning models through extensive analysis of text data to discern patterns and relationships between words, each word's numerical vector representation in the multidimensional space is highly dependent on the context in which it was trained and the spatial arrangement of the closely related words and their concepts. Techniques such as Word2Vec, GloVe (Global Vectors for word representation), and FastText, each with distinct advantages, illustrate the diversity and depth of embedding methodologies.[5–7] Furthermore, they highlight the importance of choosing the right environment in creating conceptually relevant embeddings, especially for words that are nuanced to particular applications.[8–11] Beyond theoretical applications, word embeddings have tangible implications across different aspects of medical practice, such as enhancing patient phenotyping accuracy, accelerating drug discovery, and improving the efficiency of medical literature retrieval.[12–15]

In this study, we define the "eponymic condition" as the semantic embedding of a disease's name, and the "ensemble mean" as the mean of the embeddings of all symptoms associated with that disease. We evaluate the utility of these embeddings for clinical diagnosis by testing the hypothesis that symptoms will have a closer semantic proximity to their associated disease (condition) than to other diseases (targets).

Received: October 24, 2024; **Revised:** December 7, 2024; **Editorial Decision:** December 9, 2024; **Accepted:** December 13, 2024

By comparing the distance between symptom embeddings and both the eponymic condition and the ensemble mean, we aim to determine which approach provides a more precise method for symptom-based disease classification.

To test this hypothesis, we selected 5 pediatric diseases—CHARGE syndrome, Cowden disease, POEMS syndrome, Rheumatic fever, and Tuberous sclerosis—due to their complex nature, broad spectrum of symptoms, and intricate diagnostic criteria involving both major and minor clinical signs. These sample diseases were chosen as highly challenging edge cases to rigorously evaluate the performance of embedding-based methods and stress-test the proposed approach without bias from less diagnostically complex common diseases. Symptoms were classified based on their semantic distance from both the eponymic condition (the disease name embedding) and the ensemble mean of the condition (the mean embedding of its symptoms), using Euclidean and Chebyshev distance metrics to evaluate classification accuracy.

To the best of our knowledge this approach has not been conducted in the pediatric disease space. This novel application may further advance our understanding of the potential use of embedding vectors in medical diagnosis, as well as any unnoticed pathologic or symptom clustering amongst diseases.

Methodology

Data collection and embedding generation

Symptom data for the selected diseases—CHARGE syndrome, Cowden disease, POEMS syndrome, Rheumatic fever, and Tuberous sclerosis—were collected from the widely accepted diagnostic criteria by the established sources published in PubMed.[16–20] The symptom list for Rheumatic fever was specifically analyzed for "low prevalence population" due to the simplicity of the diagnostic criteria as compared to the "high prevalence population."

To explore the semantic proximity between diseases and their symptoms, we employed the Ada-002 embedding model developed by OpenAI, which generates vector representations with a length of 1536 dimensions.[21] This model generates vector representations that encapsulate the semantic nuances of words, thereby translating textual descriptions of symptoms and diseases into a quantifiable, multidimensional vector space. To the best of our knowledge, Ada-002 was not specifically trained for medical applications, nor was its training limited to the medical domain. Our focus was particularly on embeddings of the pediatric diseases mentioned above, alongside their relevant symptoms.

Analysis of embeddings

Geometry of embeddings

The symptom embeddings in this study are represented as unit vectors in an n-dimensional space, where n in this case is 1536. Figure 1 shows the eponymic embedding vector for "CHARGE syndrome" along with the ensemble mean of all 63 symptoms for this condition. Notably, characteristic peaks at dimensions 195, 955, and 1121 are consistently observed across all eponymic conditions and their associated symptoms. The symptoms of each disease are thus mapped into a group of points on a unit $(n-1)$-sphere with the center at the mean of the symptom embeddings, as defined in eqn (4) in the Supplemental Material.

Given the spherical geometry of the embeddings, the methods of directional statistics could be employed by expressing the embeddings as $(n-1)$-dimensioinal spherical coodinates and calculating directional statistics using directional distributions such as the von Mises distribution.[22] However, as seen in the methodology sections below, this work requires only the directional or circular mean, which is derived from the mean of the Cartesian coordinates normalized to unit length, and therefore, spherical coordinates are not required in this work. For example, the directional mean of the embeddings of all 63 symptoms of CHARGE syndrome is demonstrated in Figure 1. In this study, all references to the "mean" of embeddings pertain to this directional mean. Distances calculated between the embeddings are chordal, rather than spherical. For instance, the Euclidean distance between the embeddings in Figure 1 is 0.527, with an angular distance of 0.533 radians (30.5°); given that these embeddings lie on a unit sphere, the spherical distance is also 0.533.These distances are calculated as follows, along with other distance metrics.

Distance metrics

The semantic proximity between the generated embeddings of diseases and symptoms was quantified using various distance measurement techniques, each chosen for its relevance to the nature of the data and specific analytical goals. These distance metrics were applied to the symptom embeddings generated by the Ada-002 model to evaluate the semantic proximity of diseases and symptoms. For our symptom-disease distance measurement, the methodology was based on the following approach:

Let B_1 and B_2 be embedding unit vectors, where B_1 represents the first vector and B_2 the second. These vectors are rows of numerical values that encode semantic relationships in the multidimensional space. The term B_2^{T} refers to the transpose of B_2, which is used in calculations to determine the alignment or similarity between the 2 vectors. The angle between these vectors, denoted as θ, reflects their directional relationship. A smaller angle indicates higher similarity.

The Euclidean distance, a widely used metric in vector space analysis, measures the straight-line distance (d_{E}) between 2 points in the multidimensional space. This metric was chosen for its simplicity and effectiveness in providing an overall assessment of similarity by considering all dimensions of the embedding vectors:

$$d_E = \sqrt{2(1 - \boldsymbol{B}_1\boldsymbol{B}_2^T)} \tag{1}$$

The Chebyshev distance, in contrast, focuses on the largest single-dimensional difference between 2 points. This metric highlights the maximum variation (d_{v}) across dimensions:

$$d_V = \max(|\boldsymbol{B}_1 - \boldsymbol{B}_2|) \tag{2}$$

In addition to these primary metrics, other metrics such as cosine similarity, taxicab (Manhattan) distance, and spherical distance were also considered. Cosine similarity evaluates the alignment of the 2 vectors by measuring the cosine of the angle θ, focusing on the directional relationship regardless of vector magnitude. Taxicab distance calculates the sum of the absolute differences across all dimensions, effectively measuring the cumulative "path" between 2 points. Spherical distance measures the angle θ directly between 2 vectors on the surface of a unit sphere, providing a geometric interpretation of their proximity (Supplemental Material).

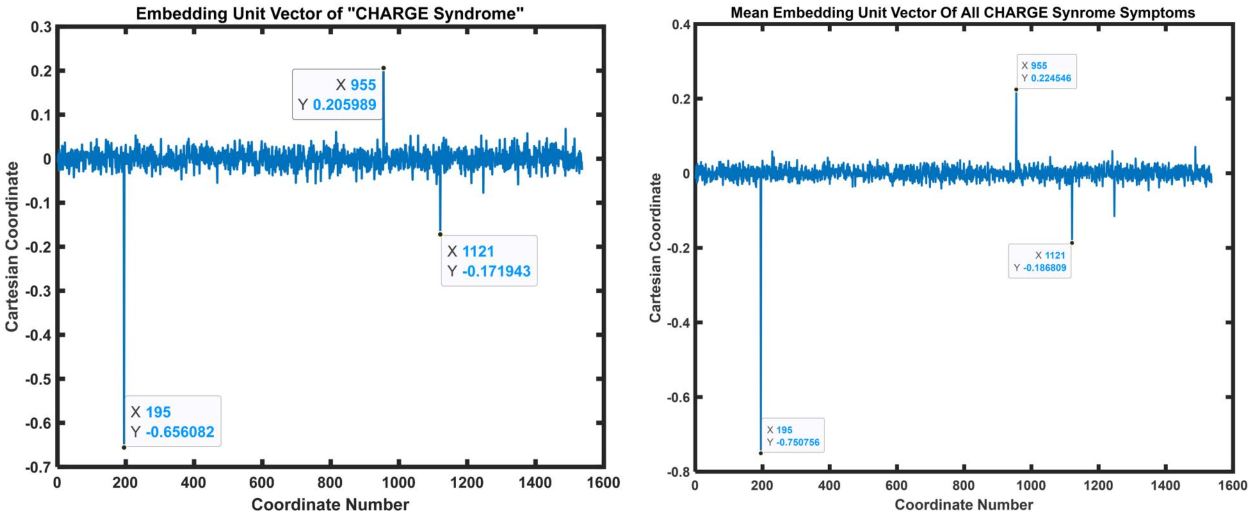

Figure 1. The Euclidean distance between the embeddings is 0.527, and the angle between them is 0.533 radians (30.5°). Since the embeddings lie on a unit sphere, the spherical distance is also 0.533.

The additional metrics, such as cosine, taxicab, and spherical distances, are simple monotomic functions of the Euclidian distance and often exhibit similar behaviors. Therefore, since Chebyshev distance is the only one of the aforementioned metrics that is not closely related to the Euclidean distance, Chebyshev and Euclidean distances were chosen to analyze our dataset.

Directional mean

To represent the collective profile of a disease, we calculated the directional mean of its associated symptom embeddings.[22] This involves averaging the embeddings of all symptoms, represented as individual vectors like B_1, and normalizing the result. The directional mean serves as a central point representing the "average" semantic position of the disease's symptom profile in the embedding space (Supplemental Material).

Kernel density distribution

To better understand the spatial arrangement of symptom embeddings, we used kernel density estimation to quantify the distribution of distances between individual symptom embeddings and the disease's directional mean. Kernel density functions smooth these distances to create a continuous curve, which helps visualize the clustering of symptom embeddings around their disease centroid.[23]

For this analysis, we used a Gaussian kernel, a common choice for smoothing, which ensures that the density function integrates to 1. The kernel density estimation was performed using MATLAB, with parameters optimized for positive support (ensuring no negative values in the computed distances). This allowed us to identify patterns in the proximity of symptoms to their associated diseases and evaluate the effectiveness of the embeddings.

Results

Classification of symptoms by semantic distance

Classification of symptoms by distance from eponymic and ensemble mean vectors

Our hypothesis is that the classification by semantic distance from the diseases' eponymic embedding vectors mirrors the clinical classification of symptoms by disease. This is tested by letting $B_k,\ \text{k}=1,\cdots,5$ be the 5 eponymic embedding vectors for each of the conditions. For each symptom, 5 distances are calculated from the B_k and the symptom is classified with the condition with the minimum semantic distance. Alternatively, we define $M_k,\ \text{k}=1,\ \cdots,\ 5$ as the ensemble means of the symptoms of each of the 5 conditions, as given in eqn (6). Similarly, we classify each symptom by the condition with the minimum distance to the corresponding M_k.

Table 1 presents the results of symptom classification using the Euclidean and Chebyshev distance metrics. The classification is considered correct when the semantic classification matches the clinical classification.

Calculating distances from the eponymic embeddings using Euclidian technique results in significant classification accuracy only for symptoms of Rheumatic fever and Tuberous sclerosis, while using the Chebyshev distance metric results in generally erratic and poor results (Table 1). However, using the ensemble means to compute distances significantly improves the classification accuracy, reducing the number of misclassified symptoms to 10 out of 152. Table 2 provides a detailed breakdown of the 10 symptoms misclassified when using the Euclidean distance to the ensemble mean of their target condition's symptom embedding vectors.

Major and minor symptoms

Categorizing symptoms into major and minor criteria enables a more granular examination of the potential correlation between semantic proximity and clinical relevance as diagnostic criteria. Major symptoms, considered to have higher diagnostic importance, are expected to be closer to the embedding point of their respective condition name. As shown in Table 3, this pattern holds true for all conditions except Cowden disease when applying the Chebyshev distance. Median values were chosen for this table to mitigate the effects of skewness or multimodality in distributions with low sample counts.

Distribution of semantic distances

Distribution of symptom distances from eponymic vectors

We calculated the distances from the eponymic embedding vector of each disease to its corresponding ensemble of associated symptoms. For example, Figure 2 shows the histogram

Table 1. Symptoms correctly classified by Euclidean and Chebyshev distance from the eponymic embedding vector and ensemble mean of the condition symptom embedding vectors.

Condition	Number of symptoms	Correctly diagnosed by eponymic embedding vector (Euclidean)	Correctly diagnosed by ensemble mean embedding vector (Euclidean)	Correctly diagnosed by eponymic embedding vector (Chebyshev)	Correctly diagnosed by ensemble mean embedding vector (Chebyshev)
CHARGE syndrome	63	2 (3%)	59 (94%)	9 (14%)	2 (3%)
Cowden disease	20	0 (0%)	16 (80%)	0 (0%)	20 (100%)
POEMS syndrome	38	1 (3%)	37 (97%)	3 (8%)	1 (3%)
Rheumatic fever	13	12 (92%)	12 (92%)	7 (54%)	0 (0%)
Tuberous sclerosis	18	18 (100%)	18 (100%)	12 (67%)	0 (0%)

Table 2. Symptoms misclassified by Euclidean distance from the ensemble mean of the condition symptom embedding vectors.

Putative condition	Incorrectly classified as	Symptom
CHARGE syndrome	POEMS syndrome	"Borderline growth hormone (GH) stimulation tests"
CHARGE syndrome	POEMS syndrome	"Growth deficiency"
CHARGE syndrome	POEMS syndrome	"Osteoporosis"
CHARGE syndrome	Rheumatic fever	"Chronic serous otitis"
Cowden disease	Tuberous sclerosis	"Fibromas"
Cowden disease	Tuberous sclerosis	"Lipomas"
Cowden disease	CHARGE syndrome	"Macrocephaly"
Cowden disease	CHARGE syndrome	"Mental retardation"
POEMS syndrome	Tuberous sclerosis	"Glomeruloid hemangiomata"
Rheumatic fever	Tuberous sclerosis	" Subcutaneous nodules"

Table 3. Distance between the eponymic embedding vectors of each condition and the ensemble of its major and minor diagnostic criteria.

Condition	Class	Group count	Median Euclidean	Median Chebyshev
CHARGE syndrome	Major	31	0.683	0.0625
CHARGE syndrome	Minor	32	0.694	0.0627
Cowden disease	Major	4	0.646	0.0646
Cowden disease	Minor	16	0.682	0.0643
POEMS syndrome	Major	6	0.661	0.0562
POEMS syndrome	Minor	25	0.706	0.0670
Rheumatic fever	Major	8	0.585	0.0525
Rheumatic fever	Minor	5	0.646	0.0672
Tuberous sclerosis	Major	11	0.585	0.0533
Tuberous sclerosis	Minor	8	0.632	0.0583

of the 63 Euclidean distances between the embeddings of CHARGE syndrome symptoms and the eponymic embedding vector for "CHARGE syndrome." The unusually small distance observed corresponds to the symptom "Distinctive CHARGE facies," which shares a term directly with the name of the condition, explaining its close proximity.

The solid blue line in Figure 2 represents the calculated kernel density function of the Euclidean distances between the ensemble of CHARGE syndrome symptom embeddings and the eponymic embedding vector of "CHARGE syndrome" as computed using eqns (7) and (8). Figure 3 illustrates the distribution of distances between the ensemble of CHARGE syndrome symptoms and all eponymic embedding vectors.

Similarly, we calculated the kernel density function for the ensemble of symptoms of all 5 conditions and the 5 eponymic embeddings vectors. If the hypothesis that semantic distances correspond to clinical findings holds true, then the distances of the symptoms of a target condition should be closest to its own eponymic vector and further from the vectors of other diseases. In Figure 4, the solid green line representing the distribution of distances between the eponymic vector of Tuberous sclerosis and its symptoms indicates that most symptoms are indeed closer to Tuberous sclerosis than to any other eponymic condition. However, this pattern does not hold for CHARGE syndrome, as shown in Figure 2. The remaining figures for Cowden disease, POEMS syndrome, and Rheumatic fever can be found in the Supplemental Material. These figures show that Cowden disease and POEMS syndrome resemble the CHARGE syndrome plot, while the plot for Rheumatic fever is similar to the Tuberous sclerosis plot.

Symptom distances to ensemble means of condition symptoms

We reasoned that the average of the symptom embeddings would offer a more accurate representation than the eponymic embedding. To test this hypostheis, we replaced the eponymic embedding vector with directional mean of all the conditions' ensemble of symptom embeddings. In this approach, a condition is defined not by its name but by its constellation of symptoms. The resulting kernel distribution plots for CHARGE syndrome and Tuberous sclerosis are seen in Figures 5 and 6, respectively. Similarly, plots for Cowden disease, POEMS syndrome, and Rheumatic fever are found in the Supplemental Material.

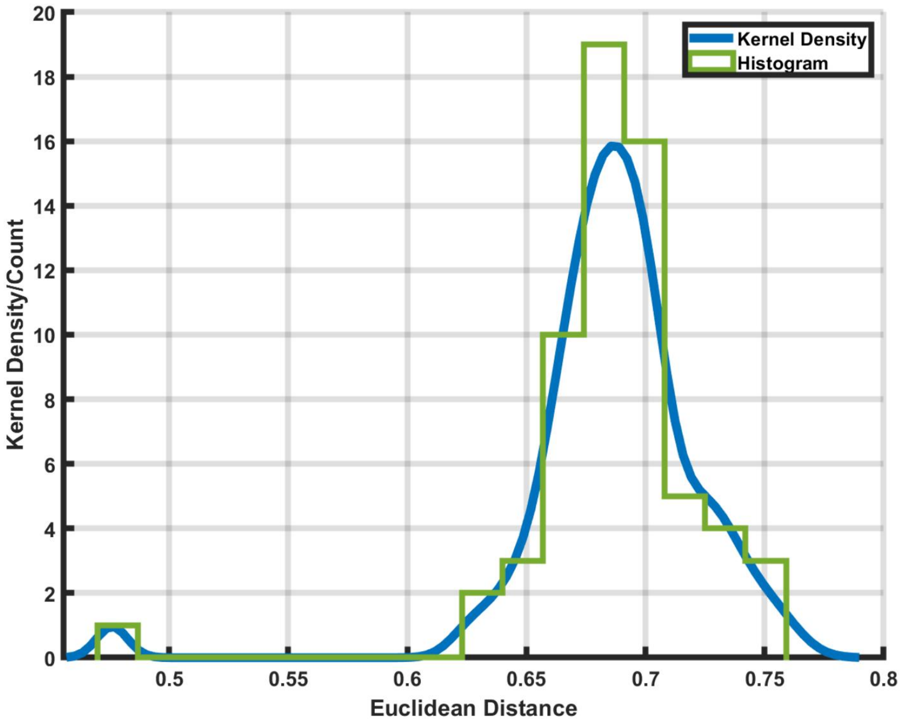

Figure 2. Distribution of distances from the ensemble of symptoms for CHARGE syndrome to the eponymic embedding vector of CHARGE syndrome. The smoothing parameter for the kernel density is h = 0.0164.

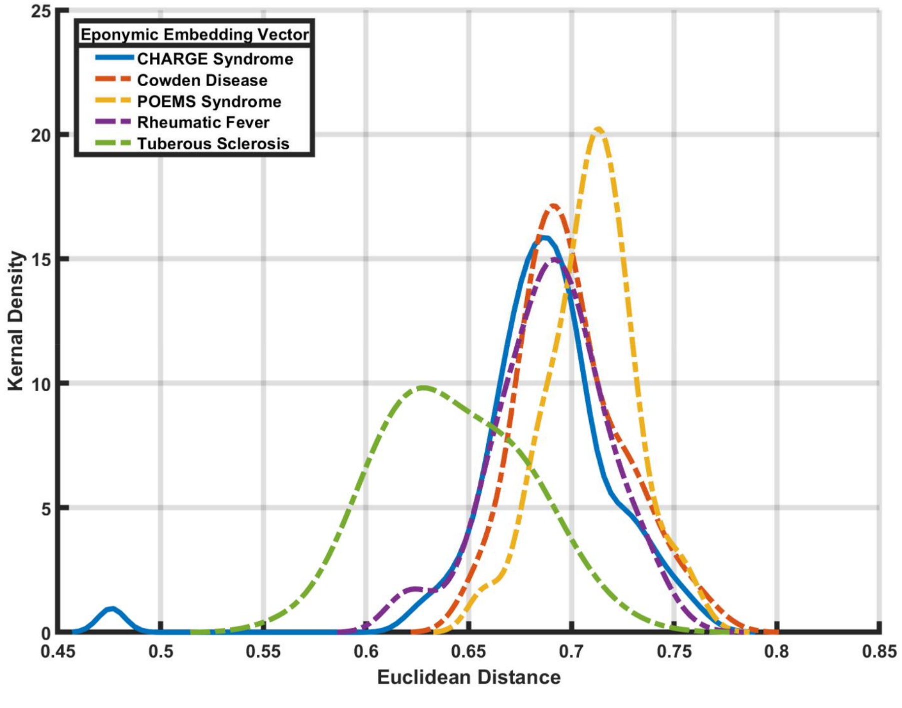

Figure 3. Distribution of distances for the ensemble of CHARGE syndrome symptoms from eponymic embedding vectors.

Discussion

This study demonstrates the potential of large language model (LLM) embeddings for clinical diagnosis by quantifying the semantic proximity between medical terms. We explored the utility of 2 approaches: classifying symptoms by their distance from the eponymic embedding of a disease name ("eponymic condition") versus the mean embedding of all symptoms associated with a disease ("ensemble mean").

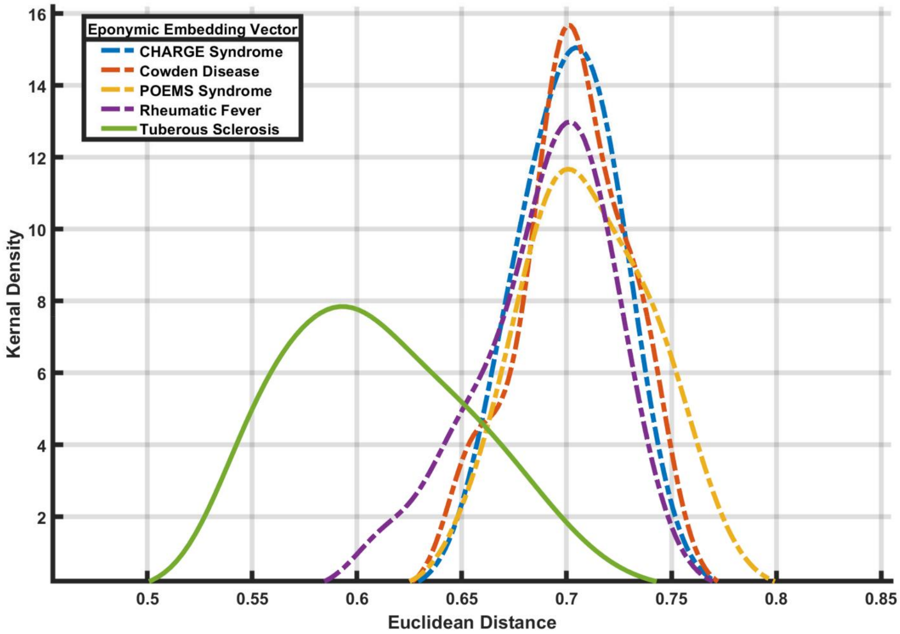

Figure 4. Distribution of distances for the ensemble of tuberous sclerosis symptoms from eponymic embedding vectors.

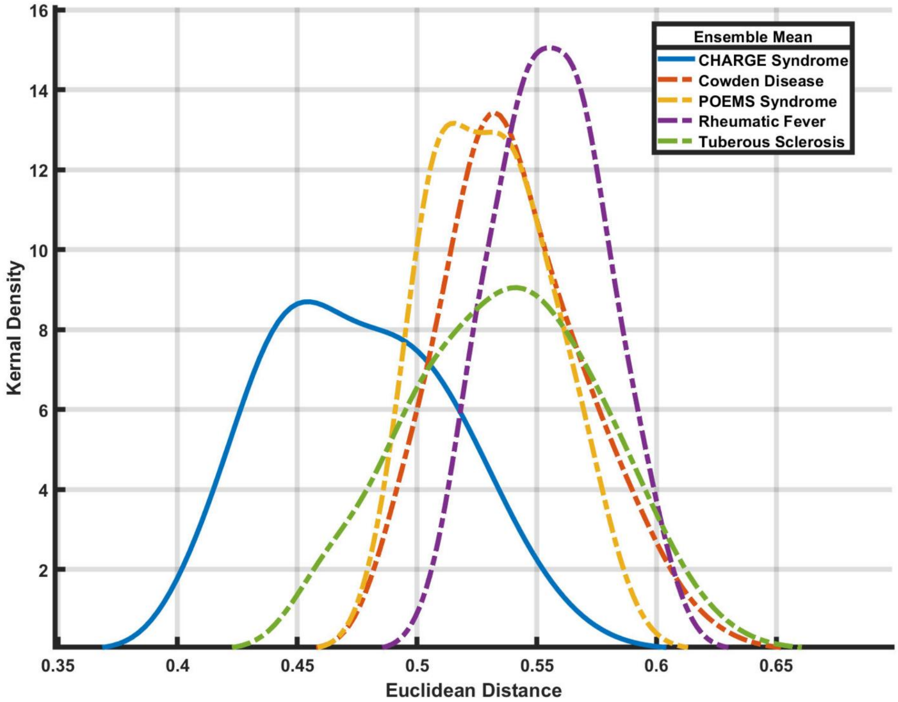

Figure 5. Distribution of distances from all symptoms of CHARGE syndrome to ensemble mean embedding vectors.

Our findings highlight both the strengths and limitations of these methods and suggest future directions for improving their clinical applicability.

The result of this study suggests that the ensemble mean of symptom embeddings provides a more accurate basis for symptom classification than the eponymic embeddings of disease names. Using the Euclidean distance metric, the ensemble mean approach significantly outperformed the eponymic approach, achieving a range of 80%-100% accuracy. This suggests that aggregating the semantic representations of symptoms into a composite vector of its symptoms better embodies the unique characteristics of a disease than relying solely on the disease name embedding. Moreover, the relatively closer proximity of the median Euclidean distance of major symptoms' aggregates to their target condition's eponymic embedding, compared to the distances of minor symptoms' embeddings, demonstrates the potential of semantic embeddings to align with real-life clinical concepts. In contrast, the Chebyshev distance metric, which focuses on maximum deviation along any dimension, performed poorly and inconsistently throughout most measurements. Chebyshev's sensitivity to outliers likely reduces its utility for capturing complex, multidimensional relationships in

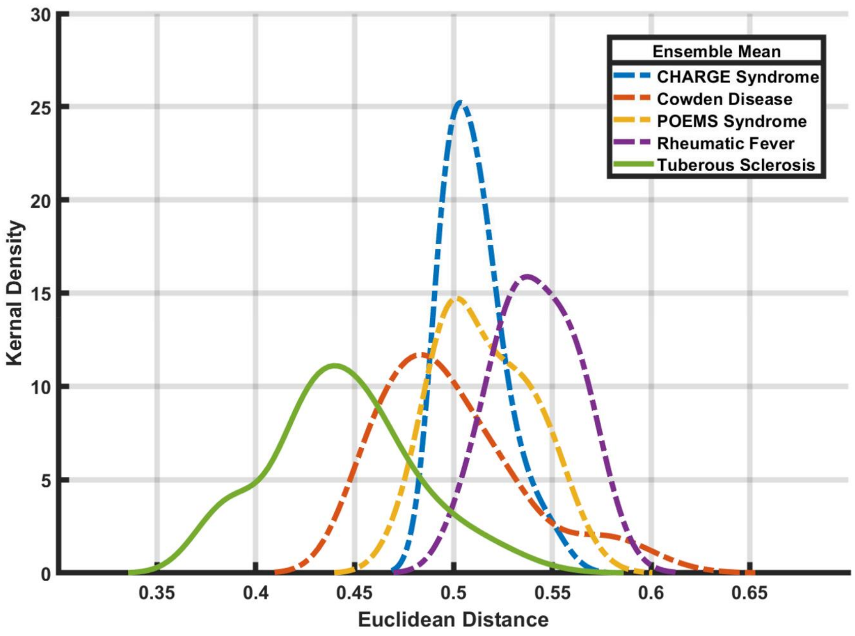

Figure 6. Distribution of distances from all symptoms of tuberous sclerosis to ensemble mean embedding vectors.

medical data, where symptoms often vary widely in presentation and severity.

The superior performance of the ensemble mean approach highlights the importance of context-rich representations in clinical settings. By capturing the collective profile of symptoms, this method provides a nuanced view of disease presentation that is highly relevant for differential diagnosis, particularly where overlapping clinical features create challenges. However, the misclassifications observed in Table 2 provide further evidence of the complexities involved in using embeddings for diagnosis. These errors suggest that embeddings may capture secondary tangential associations arising from clinical overlap or semantic similarity rather than direct clinical relevance. For instance, certain symptoms associated with CHARGE syndrome, such as "Borderline growth hormone (GH) stimulation tests," "Growth deficiency," and "Osteoporosis," were misclassified under POEMS syndrome due to the semantic similarities with endocrine and growth-related abnormalities seen in both diseases.[24]

Similarly, "Chronic serous otitis" was misclassified as "Rheumatic fever," likely due to the shared involvement of streptococcal infections, which can cause "otitis media" (*Streptococcus pneumoniae)*, and if left untreated as pharyngeal infection, it may lead to Rheumatic fever (*Streptococcus pyogenes)*.

These examples underscore the need for more specialized training of LLMs using medical data to better differentiate between diseases with superficially similar, but clinically distinct features. Embeddings that are more attuned to the specific language and complexities of medical discourse could enhance accuracy by recognizing subtle differences in symptomatology, comorbidities, and disease progression, which are critical for precise diagnosis and choosing the most effective treatment.

This study highlights the importance of further exploring the diagnostic specificity and contextual relevance of embeddings by applying this approach across a broader spectrum of diseases, including those with overlapping symptoms and varying levels of diagnostic complexity, to evaluate its generalizability and limitations. To improve the specificity and clinical relevance of word embeddings for medical diagnosis, future models may benefit from being predominantly trained on comprehensive and diverse medical texts, such as electronic health records (EHRs), clinical guidelines, and case studies, to better capture the subtle nuances and specificities of medical terminology. This may minimize nonspecific associations and improve the ability of models to identify true diagnostic clues. Furthermore, combining LLM-based embeddings with other data sources, such as genetic, imaging, and laboratory data, could enhance the utility and diagnostic accuracy of the models by providing a more holistic view of disease presentation.

Additionally, developing new distance metrics that incorporate clinical context, such as weighted distances that prioritize diagnostically relevant symptoms over less specific ones, could further refine embedding-based classification methods. For example, metrics could be designed to give greater importance to major diagnostic criteria while minimizing the influence of common, nonspecific symptoms.

Regardless of the methodology used, validation of these models across a wide range of common and rare diseases, clinical scenarios, and populations is essential to avoid biases that may disseminate from overfitting for a sub-group of diseases, or regional and socioeconomical nuances of the training datasets. Real-world testing in clinical environments, such as hospitals or outpatient clinics, will be crucial for understanding the practical utility and limitations of these models. Moreover, since retraining an entire LLM with continuing medical advancement and changing literature is both energy intense and laborious, the use of techniques such as the ensemble mean can make diagnostic models more easily adaptable to any new updates or refinements.

Limitations

This study's findings must be interpreted within the context of several limitations. The analysis initially focused on a limited number of pediatric diseases with rarely overlapping symptoms, which may not accurately represent the broader range of conditions encountered in clinical practice or the complexity involved in differentiating between 2 distinct diseases with many overlapping symptoms. Pediatric conditions often have distinct symptom profiles and diagnostic criteria that may not directly translate to adult or geriatric populations, thus limiting the generalizability of our results.

Furthermore, the study primarily utilized Euclidean and Chebyshev distance metrics, which may not encompass the full spectrum of distance metrics that could offer greater specificity and accuracy. Future research should explore a broader range of distance models and embedding techniques to enhance the diagnostic utility of semantic proximity.

Another critical limitation is the reliance on embeddings from LLMs not specifically trained in a medical context. The use of alternate models may offer a varying result.

Conclusion

Word embeddings from LLMs offer promising avenues for enhancing medical diagnosis, but their clinical utility requires further refinement to achieve optimal performance. While the ensemble mean approach shows improved classification accuracy over eponymic embeddings, significant challenges remain, particularly regarding specificity and contextual relevance. Addressing these challenges will require targeted model training on medical data, development of hybrid approaches, and creation of context-aware metrics that reflect the complexity of clinical decision-making.

Future research efforts should focus on refining these models to enhance their diagnostic accuracy and clinical applicability. By leveraging specialized training, integrating diverse data sources, and employing dynamic learning techniques, LLM-based embeddings can evolve to meet the unique demands of clinical practice. As these tools mature, they have the potential to transform diagnostics, personalized medicine, and other areas of healthcare by providing a robust computational framework for interpreting complex medical data. This evolution could ultimately lead to more precise, efficient, and individualized patient care.

Author contributions

All authors were involved in all aspects of this research, including the design, data processing, and writing of the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Shickel B, Tighe PJ, Bihorac A, et al. Deep EHR: a survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2018;22:1589-1604.
2. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317-1318.
3. Zhang Y, Jatowt A, Tanaka K. Towards understanding word embeddings: automatically explaining similarity of terms. In: *2016 IEEE International Conference on Big Data*.
4. Si Y, Wang J, Xu H, et al. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc*. 2019;26:1297-1304. arXiv:1902.08691.
5. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv, preprint arXiv:1301.3781, preprint: not peer reviewed.
6. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014:1532-1543.
7. Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information. *TACL*. 2017;5:135-146.
8. Tekir S, Sezerer E. A survey on neural word embeddings. arXiv, preprint arXiv2110.01804, preprint: not peer reviewed.
9. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234-1240.
10. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv, preprint arXiv:1904.03323, preprint: not peer reviewed.
11. Beltagy I, Lo K, Cohan A. SciBERT: a pre-trained language model for scientific text. arXiv, preprint arXiv:1903.10676, preprint: not peer reviewed.
12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.
13. Park K. A review of computational drug repurposing. *Transl Clin Pharmacol*. 2019;27:59-63.
14. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20.
15. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13:e0192360.
16. Blake KD, Davenport SL, Hall BD, et al. CHARGE association: an update and review for the primary pediatrician. *Clin Pediatr (Phila)*. 1998;37:159-173.
17. Pilarski R, Burt R, Kohlman W, et al. Cowden syndrome and the PTEN hamartoma tumor syndrome: systematic review and revised diagnostic criteria. *J Natl Cancer Inst*. 2013;105:1607-1616.
18. Dispenzieri A. POEMS syndrome. *Blood Rev*. 2007;21:285-299.
19. Gewitz MH, Baltimore RS, Tani LY, American Heart Association Committee on Rheumatic Fever, Endocarditis, and Kawasaki Disease of the Council on Cardiovascular Disease in the Young, et al. Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: a scientific statement from the American Heart Association. *Circulation*. 2015;131:1806-1818.

20. Northrup H, Aronow ME, Bebin EM, International Tuberous Sclerosis Complex Consensus Group, et al. Updated international tuberous sclerosis complex diagnostic criteria and surveillance and management recommendations. *Pediatr Neurol*. 2021;123:50-66.
21. OpenAI. *Ada-002 embedding model documentation*. Accessed October 3, 2024. https://platform.openai.com/docs/models
22. Mardia KV, Jupp PE. *Directional Statistics*. Wiley Series in Probability and Statistics. 2nd ed. John Wiley & Sons; 2009.
23. Härdle W, Müller M, Sperlich S, et al. *Nonparametric and Semiparametric Methods*. Springer-Verlag; 2004.
24. Wang Y, Liu Y, Huang X, et al. POEMS syndrome characterized by bone lesions: a case report. *Medicine (Baltimore)*. 2023;102:e36678.

Journal of the American Medical Informatics Association, 2025, 00, 1–9
https://doi.org/10.1093/jamia/ocae314
Research and Applications