CME LIVE OPEN ACCESS

Journal of
Clinical and Medical Engineering Live

Research Article

Medical Diagnosis Coding Automation: Similarity Search vs. Generative Ai

Vanessa Klotzman[1&2*]

[1]*Department of Informatics, University of California, Irvine, 6210 Donald Bren Hall, Irvine, 92697-3425, CA, United States.*

[2]*Research Institute, Children's Hospital of Orange County, 1201 W. La Veta Ave., Orange County, 92868, CA, United States.*

Corresponding Author: Vanessa Klotzman, Department of Informatics, University of California, Irvine, 6210 Donald Bren Hall, Irvine, 92697-3425, CA, United States. Research Institute, Children's Hospital of Orange County, 1201 W. La Veta Ave., Orange County, 92868, CA, United States.

Received: 2024 Jul 09 **Accepted:** 2024 Jul 29 **Published:** 2024 Aug 02

Abstract

Objective: This study aims to predict ICD-10-CM codes for medical diagnoses from short diagnosis descriptions and compare two distinct.

Approaches: similarity search and using a generative model with few-shot learning.

Materials and Methods: The text-embedding-ada-002 model was used to embed textual descriptions of 2023 ICD-10-CM diagnosis codes, provided by the Centers provided for Medicare & Medicaid Services. GPT-4 used few-shot learning. Both models underwent performance testing on 666 data points from the eICU Collaborative Research Database.

Results: The text-embedding-ada-002 model successfully identified the relevant code from a set of similar codes 80% of the time, while GPT-4 achieved a 50 % accuracy in predicting the correct code.

Discussion: The work implies that text-embedding-ada-002 could automate medical coding better than GPT-4, highlighting potential limitations of generative language models for complicated tasks like this.

Conclusion: The research shows that text-embedding-ada-002 outperforms GPT4 in medical coding, highlighting embedding models' usefulness in the domain of medical coding.

Keywords: Embeddings, Large Language Models, ICD Codes and Automated Medical Coding.

1. Introduction

The International Classification of Disease (ICD), established by the World Health Organization (WHO) is the universally recognized and standardized system for medical coding worldwide. It provides a comprehensive framework for categorizing diseases, health conditions, and related information, facilitating accurate and consistent documentation, data sharing, and research across the global healthcare community. It is employed by healthcare providers worldwide to categorize diseases and conditions. Medical coding involves the assignment of ICD codes like ICD-10-CM codes to classify diagnoses and reasons for visits in all healthcare settings, is essential for guiding clinical decisions, tracking diseases, and impacting healthcare financing [1, 2]. Medical coding is traditionally manual, with coders translating physicians' notes into the appropriate ICD codes while adhering to complex guidelines. In this process, highly trained medical coders assign ICD (International Classification of Diseases) codes to patient encounters based on the information found in clinicians' notes, however, manual ICD coding is time-consuming and error-prone, making the quality and productivity of coding a matter of concern in practice. The process is error-prone [3–5] due to the complexity of medical language and coding guidelines. Coders often need help with subtle differences between disease subtypes, leading to misclassification. Physicians' use of abbreviations and synonyms which adds to the ambiguity [6]. Making this a non-trivial task for humans. Furthermore, inexperienced coders may incorrectly assign separate codes to related diagnoses, a problem called unbundling, which can result in costly mistakes [7]. These coding inaccuracies have substantial financial implications, contributing to an estimated annual expenditure of $25 billion in the United States, as reported by Lang et al. [8] Farkas et al. [9]. With recent AI technologies (e.g., NLP), automated medical coding has the potential to support clinical coders better.

Automated Medical Coding (AMC) is the idea that artificial intelligence can automate clinical coding. In recent years, there has been a significant increase in AMCrelated work [10–17] through deep learning. Although research in this field has grown, this problem is far from being solved [18, 19]. For instance, automated coding remains a complex problem because extracting knowledge from patients' clinical records is challenging. These records are not uniformly structured, the medical field's terminologies can be complicated for non-professionals to comprehend, and physicians often have different ways of describing symptoms, leading to various descriptions for the same disease.

Embeddings in Natural Language Processing (NLP) represent words as realvalued vectors [20]. These vectors can capture the meaning of words in such a way that words closer together in the vector space are expected to have similar meanings. In clinical NLP, embeddings are helpful for analyzing medical data and texts, aiding decision-making and research [21]. The use of word embeddings in Automated Medical Coding (AMC) systems is increasingly being explored as it has the potential to bridge the gap between the informal language of medical diagnoses and the formal language of ICD code descriptions [22-27]. For instance, CAIC uses cross-textual attention to match parts of medical notes with ICD codes [15]. While GatedCNN-NCI creates a network linking every aspect of medical notes to ICD codes [19]. BiCapsNetLE integrates ICD code descriptions into word embeddings of clinical notes, enhancing alignment [28]. DLAC employs a description-based label attention mechanism, focusing on the correlation between the descriptions of ICD codes and the features of medical notes [29]. ICDBigBird uses a Graph Convolutional Network (GCN) and ehances the ICD code emebddings by using their relational structure.

Even though, there is a growing body of work for utilizing embeddings in clinical coding, there has been a growing interest of what a Large Language Model (LLM) can do in the health sector due to their ability of understanding, generating, and predicting new content.

As the interest in Large Language Models (LLMs) continues to grow in the health sector, as evidenced by multiple recent studies studies, our objective is to compare the effectiveness of two distinct approaches to predict ICD-10-CM codes accurately [30-34]. We will compare the effectiveness of similarity search, for which we will be using text-embedding-ada-002, and an LLM, in which we will be using GPT-4 from OpenAI to predict ICD-10-CM codes [35].

2. Materials and Methods

Data Collection: We utilized the diagnosis strings (patient diagnoses) from the eICU Collaborative Research Database, which contains data from different critical care units (CCUs) across the United States from patients who were admitted between 2014 and 2015 [36]. We selected a subset of 666 patients from the total dataset of 2,710,672 patients. This sample size represents a 99% confidence level with a 5% margin of error [37]. We utilize each patient's current diagnoses from the data we collected, which comprise of the diagnosis string and the corresponding ICD-10 CM codes. The diagnosis strings will serve as inputs to the models, with the ICD-10-CM codes as the outputs. The ICD-10-CM codes will be used for comparison to assess the model's accuracy in prediction. The diagnosis strings in the eICU database are organized in a tiered system. For example, "neurologic—trauma - CNS—intracranial injury—with subarachnoid hemorrhage" shows this: it starts with a general category "neurologic", goes into a more specific "trauma - CNS", then to "intracranial injury", and ends with a detailed aspect "with subarachnoid hemorrhage". Each part of the string represents a deeper level of diagnosis detail. Table 1, shows sample diagnosis strings and their corresponding ICD-10 CM codes from the dataset we are using.

Table 1: eICU Sample Data: Diagnosis & ICD-10 Code.

Diagnosis	ICD-10 CM Code
burns/trauma|dermatology|cellulitis	L03.90
burns/trauma|trauma - chest|lung trauma	S27.30
hematology|coagulation disorders|DVT	I80.9

Model Selection: We utilize OpenAI's text-embedding-ada-002 model, as it surpasses previous models in text search and text similarity from OpenAI [1]. We evaluate the effectiveness of the embedding model relative to the latest version of GPT-4, which we operated using the Microsoft OpenAI Azure Service. Our selection of only the text-embedding-ada-002 model and GPT-4 was due to the constraints set by the terms and conditions of using the PhysioNet dataset [1].

Text-Embedding-Ada-002: In this work, we utilized the text-embedding-ada-002 model to embed the textual descriptions of 2023 ICD-10-CM diagnosis codes, as provided by the Centers for Medicare & Medicaid Services source [3]. After generating these embeddings, our primary objective was to evaluate their performance. To do so, we used the dataset of diagnosis strings obtained from the eICU dataset.

To assess the accuracy of matching medical diagnoses (diagnosis strings) with their respective ICD-10-CM codes, we inputted these diagnosis strings into the textembedding-ada-002 model. Our objective was to determine if this single model could accurately return the closest ICD-10-CM code based on the ICD-10 description. Additionally, the model returned the top four ICD-10-CM codes for each medical condition. We selected four as the default value for similarity searches by vector, because this is specified as the standard setting in the LangChain documentation [4]. The workflow of

how the embeddings function is clearly illustrated in Figure 1 The embeddings are generated and then stored in the vector database. These embeddings correspond to the textual descriptions of the 2023 ICD-10-CM diagnosis codes. When a user submits a query as a medical diagnosis (referred to as the diagnosis string), the system searches the database for embeddings similar to the embedding of the query. Finally, the system retrieves the closest ICD-10-CM codes based on the similarity between embeddings, providing relevant matches for the medical diagnosis.

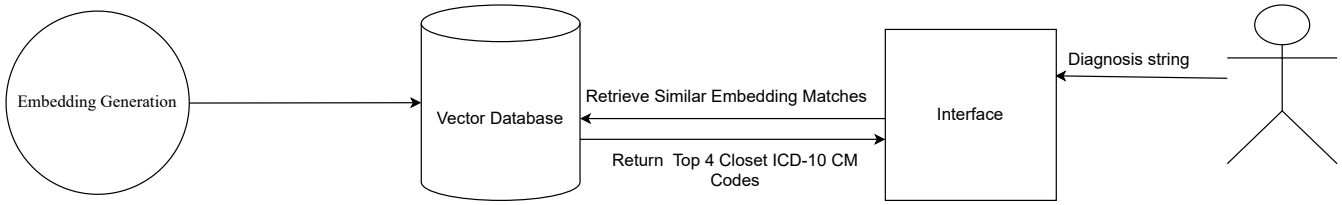

Figure 1: Visualizing Automated ICD-10 Code Prediction Process: Streamlining Medical Coding.

GPT-4: We prompted GPT-4 with few-shot prompting to assess its capability in medical coding. Few shots prompting was selected because large language models have notable zero-shot abilities, but they tend to perform poorly in complex tasks when using zero-shot settings [38]. Few-shot prompting serves as a method to enable in-context learning, where demonstrations in the prompt help direct the model toward better performance. Figure 2 contains the prompt we used. The examples for the prompt were acquired by clustering the diagnosis strings. We used K-means clustering to group our diagnosis strings and found that 8 clusters worked best.

The ideal number of eight clusters was determined using the elbow method, which evaluates the withincluster sum of squares across a range of 1 to 20 possible clusters; the 'elbow' point, where there is a significant decrease in within-cluster dissimilarity, indicates the most suitable number of clusters. We picked a range between 1 and 20 as it is a manageable number of clusters that can be effectively interpreted and analyzed. From each cluster, we selected the diagnosis closest to the cluster's center in the vector space as the most representative example of the cluster; these representative examples were then used in the few-shot prompt.

To optimize GPT-4 for medical coding, we experimented with different temperature settings (0.1, 0.5, 0.9) on the sample of 666 diagnoses collected for this study, representing a 99% confidence interval and a 5% margin of error. Our results showed that a temperature of 0.1 was effective, as it balanced the model's creative outputs and the need for accurate, deterministic responses in medical coding.

Role: ***Medical Coder Objective:***
Your task is to accurately assign the correct ICD-10-CM code for each patient's condition based on their medical diagnosis.

Examples:
- Description: neurologic disorders of vasculature stroke
- Output: I67.8
- Description: infectious diseases systemic/other infections sepsis
- Output: A41.9
- Description: pulmonary disorders of vasculature pulmonary embolism
- Output: I26.99
- Description: burns/trauma|trauma - CNS|intracranial injury
- Output: S06.9
- Description: cardiovascular ventricular disorders congestive heart failure
- Output: I50.9
- Description: gastrointestinal GI bleeding / PUD|peptic ulcer disease
- Output: K27.9
- Description: pulmonary respiratory failure acute respiratory failure
- Output: J96.00
- Description: renal disorder of kidney acute renal failure
- Output: N17.9
- Description: ${input text} $ Output.

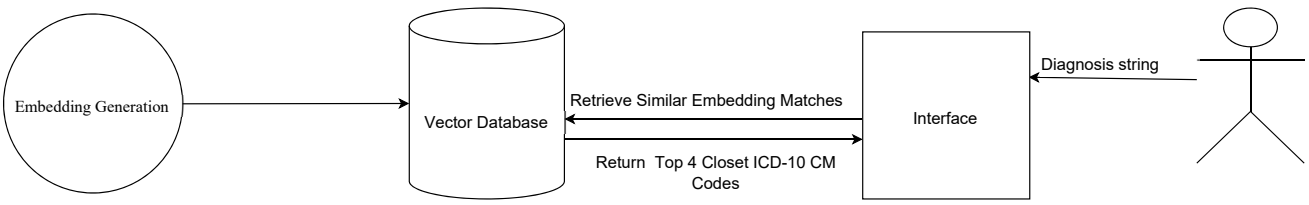

Figure 2: GPT-4 Prompt.

3. Results

The text-embedding-ada-002 model achieved an 80% accuracy rate in identifying the correct ICD-10-CM codes from the retrieved similar codes, outperforming GPT-4, which achieved a 50% accuracy rate in the same task. This suggests that embedding models, like text-embedding-ada-002, can offer improved accuracy and efficiency in medical coding.

4. Discussion

In this study, we discovered that embedding models like text-embedding-ada-002 could potentially be more effective than GPT-4, a large language model. The critical advantage of embeddings lies in their focus on the semantic similarity of words, an aspect vital for accurately matching medical diagnoses with ICD codes, as this technique allows for a more precise understanding and interpretation of medical terminology, which is crucial in medical coding. Embeddings analyze the context and meaning of words more concentratedly, leading to higher accuracy in identifying relevant codes.

Moreover, when assessing the feasibility of using embedding models like textembedding-ada-002, it becomes evident that these models align well with medical coding requirements. They offer a more focused approach, potentially assisting in accurately linking diagnoses with the correct ICD codes, which demands precision. It suggests that embedding models better fit medical coding tasks compared to more generative models like GPT-4, which handle a broader range of data.

In contrast, GPT-4 processes a wide range of data and contexts. While this versatility is helpful for general tasks, it can lead to less precision in specialized areas like medical coding, where specific terminology and accurate coding are essential. GPT-4's handling of vast information might make it more challenging to differentiate between similar medical terms and codes, potentially affecting its performance in this field.

5. Conclusion

The results indicate that embedding models like text-embedding-ada-002 appear more suitable for medical coding tasks than large language models like GPT-4. This result could be primarily due to text-embedding-ada-002's focused approach on the semantic similarity of words, which has led to an 80% accuracy in identifying ICD10-CM codes, significantly surpassing GPT-4's 50% accuracy. Embedding models like text-embedding-ada-002 are a more practical choice for medical coding due to their precision in analyzing and understanding medical terminology. On the other hand, GPT-4, although capable of broad data processing, may prove less effective in specialized fields such as medical coding, where accuracy and specific terminology are crucial. Hence, for precision-dependent tasks such as medical coding, embedding models like text-embedding-ada-002 could offer a more suitable solution than generative models like GPT-4.

References

1. Aalseth, P. (2014). *Medical Coding: What it is and how it Works.* Jones & Bartlett Publishers.
2. Borman, K. R. (2017). Medical Coding in the United States: Introduction and Historical Overview. *Principles of Coding and Reimbursement for Surgeons*, 3-11.
3. Asadi, F., Hosseini, M. A., Gomar, T., & Sabahi, A. (2022). Factors Affecting Clinical Coding Errors. *Shiraz E-Medical Journal, 23*(9).
4. Lloyd, S. S., & Rissing, J. P. (1985). Physician and coding errors in patient records. *Jama, 254*(10), 1330-1336.
5. Cipparone, C. W., Withiam-Leitch, M., Kimminau, K. S., Fox, C. H., Singh, R., et al (2015). Inaccuracy of ICD-9 codes for chronic kidney disease: a study from two practice-based research networks (PBRNs). *The Journal of the American Board of Family Medicine, 28*(5), 678-682.
6. Xie, P., & Xing, E. (2018, July). A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1066-1076).
7. Alonso, V., Santos, J. V., Pinto, M., Ferreira, J., Lema, I., et al (2020). Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. *Journal of medical systems, 44*, 1-8.
8. Lang, D. (2007). Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter, 6.*
9. Farkas, R., & Szarvas, G. (2008, April). Automatic construction of rule-based ICD-9-CM coding systems. In *BMC bioinformatics* (Vol. 9, pp. 1-9). BioMed Central.
10. Teng, F., Liu, Y., Li, T., Zhang, Y., Li, S., et al (2022). A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering, 35*(5), 4357-4375.
11. Chen, Y., Chen, H., Lu, X., Duan, H., He, S., et al (2023). Automatic ICD-10 coding: Deep semantic matching based on analogical reasoning. *Heliyon, 9*(4).
12. Moons, E., Khanna, A., Akkasi, A., & Moens, M. F. (2020). A comparison of deep learning methods for ICD coding of clinical records. *Applied Sciences, 10*(15), 5262.
13. Ramalho, A., Souza, J., & Freitas, A. (2020, June). The use of artificial intelligence for clinical coding automation: a bibliometric analysis. *In International Symposium on Distributed Computing and Artificial Intelligence* (pp. 274-283). Cham: Springer International Publishing.
14. Sonabend, A., Cai, W., Ahuja, Y., Ananthakrishnan, A., Xia, Z., et al (2020). Automated ICD coding via unsupervised knowledge integration (UNITE). *International journal of medical informatics,* 139, 104135.
15. Teng, F., Ma, Z., Chen, J., Xiao, M., & Huang, L. et al (2020). Automatic medical code assignment via deep learning approach for intelligent healthcare. *IEEE journal of biomedical and health informatics, 24*(9), 2506-2515.
16. Kaur, R., Ginige, J. A., & Obst, O. (2023). AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Systems with Applications, 213,* 118997.
17. Wang, S. M., Chang, Y. H., Kuo, L. C., Lai, F., Chen, Y. N., et al (2020). Using deep learning for automatic ICD-10 classification from free-text data. *European Journal of Biomedical Informatics, 16*(1), 1-10.

18. Yan, C., Fu, X., Liu, X., Zhang, Y., Gao, Y., et al (2022). A survey of automated International Classification of Diseases coding: development, challenges, and applications. *Intelligent Medicine, 2*(03), 161-173.
19. Ji, S., Li, X., Sun, W., Dong, H., Taalas, A., et al (2022). A unified review of deep learning for automated medical coding. *ACM Computing Surveys.*
20. Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069.*
21. Kalyan, K. S., & Sangeetha, S. (2020). SECNLP: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics, 101,* 103323.
22. Nath, N., Lee, S. H., & Lee, I. (2023). Application of specialized word embeddings and named entity and attribute recognition to the problem of unsupervised automated clinical coding. *Computers in Biology and Medicine, 165*, 107422.
23. Biseda, B., Desai, G., Lin, H., & Philip, A. (2020). Prediction of ICD codes with clinical BERT embeddings and text augmentation with label balancing using MIMIC-III. *arXiv preprint arXiv:2008.10492.*
24. Yogarajan, V., Gouk, H., Smith, T., Mayo, M., & Pfahringer, B. et al (2020). Comparing high dimensional word embeddings trained on medical text to bag-of-words for predicting medical codes. In *Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings, Part I 12* (pp. 97-108). Springer International Publishing.
25. Zhang, S., Zhang, B., Zhang, F., Sang, B., & Yang, W. et al (2022, October). Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2883-2891).
26. Steiger, E., & Kroll, L. E. (2023). Patient embeddings from diagnosis codes for health care prediction tasks: Pat2Vec machine learning framework. *JMIR AI, 2,* e40755.
27. Shi, W., Wu, J., Yang, X., Chen, N., Mien, I. H., et al (2021, August). Analyzing Code Embeddings for Coding Clinical Narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4665-4672).
28. Bao, W., Lin, H., Zhang, Y., Wang, J., & Zhang, S. et al (2021). Medical code prediction via capsule networks and ICD knowledge. *BMC Medical Informatics and Decision Making, 21,* 1-12.
29. Feucht, M., Wu, Z., Althammer, S., & Tresp, V. (2021). Description-based label attention classifier for explainable ICD-9 classification. *arXiv preprint arXiv:2109.12026.*
30. Sezgin, E., Chekeni, F., Lee, J., & Keim, S. (2023). Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. *Journal of Medical Internet Research, 25*, e49240.
31. Wagner, M. W., & Ertl-Wagner, B. B. (2023). Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, 08465371231171125.
32. Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., et al (2023). Evaluating large language models on medical evidence summarization. *NPJ digital medicine, 6*(1), 158.
33. Dong, H., Falis, M., Whiteley, W., Alex, B., Matterson, J., et al (2022). Automated clinical coding: what, why, and where we are? *NPJ digital medicine, 5*(1), 159.
34. Shah, N. H., Entwistle, D., & Pfeffer, M. A. (2023). Creation and adoption of large language models in medicine. *Jama, 330*(9), 866-869.
35. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., et al (2022). Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005.*
36. Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., et al (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data, 5*(1), 1-13.
37. Altman, D., Machin, D., Bryant, T., & Gardner, M. (Eds.). (2013). *Statistics with confidence: confidence intervals and statistical guidelines.* John Wiley & Sons.
38. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., et al (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33,* 1877-1901.