Retrieval-Augmented Generation for Extracting CHA_2DS_2-VASc Risk Factors from Unstructured Clinical Notes in Patients with Atrial Fibrillation

Philip Adejumo BS[1], Phyllis Thangaraj MD, PhD[1], Sumukh Vasisht Shankar MS[1], Lovedeep Singh Dhingra MBBS[1], Arya Aminorroaya MD, MPH[1], Rohan Khera MD, MS[1,2,3,4]

1. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT
2. Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT
3. Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT
4. Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT

Address for correspondence:
Rohan Khera, MD, MS
195 Church St, 6th Floor, New Haven, CT 06510
203-764-5885; rohan.khera@yale.edu; @rohan_khera

Abstract

Background: Assessment of stroke risk in patients with atrial fibrillation (AF) is crucial for guiding anticoagulation therapy. CHA_2DS_2-VASc is a widely used score for defining this risk, but current assessments rely on manual calculation by clinicians or approximations from structured EHR data elements. Unstructured clinical notes contain rich information that could enhance risk assessment. We developed and validated a Retrieval-Augmented Generation (RAG) approach to extract CHA_2DS_2-VASc risk factors from unstructured notes in patients with AF.

Methods: We employed a RAG architecture paired with the large language model, Llama3.1, to extract features relevant to CHA_2DS_2-VASc scores from unstructured notes. The model was deployed on a random set of 1,000 clinical notes (934 AF patients) from Yale New Haven Health System (YNHHS). To establish a gold standard, 2 clinicians manually reviewed and labeled CHA_2DS_2-VASc risk factors in a random subset of 200 notes. The CHA_2DS_2-VASc scores were calculated for each patient using structured data alone and by incorporating risk factors identified with RAG. We assessed performance across risk factors using macro-averaged area under the receiver operating characteristic (AUROC). For external validation, we utilized 100 manually labeled clinical notes from the MIMIC-IV database.

Results: The RAG model demonstrated robust performance in extracting risk factors from clinical notes. In the 1000 clinical notes, RAG identified several risk factors more frequently than structured elements, including hypertension (82.4% vs 26.2%), stroke/TIA (62.9% vs 45.5%), vascular disease (83.4% vs 56.6%), and diabetes (84.1% vs 47.2%). In the 200 expert-annotated notes, the RAG approach achieved high performance for various risk factors, with AUROCs ranging from 0.96 to 0.98 for hypertension, diabetes, and age ≥75 years. Incorporating risk factors identified by RAG increased CHA_2DS_2-VASc scores compared with using structured data alone.

Conclusion: An LLM-optimized RAG can accurately extract CHA_2DS_2-VASc risk factors from unstructured clinical notes in AF patients. This approach can enable computable risk assessment and guide appropriate anticoagulation therapy.

Background

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, affecting an estimated 33 million individuals worldwide leading to a high burden of morbidity and mortality causing 10-12% of strokes worldwide.[1] Medications such as direct oral anticoagulants (DOACs) can reduce the risk of stroke and lead to improved cardiovascular outcomes in patient with atrial fibrillation. Per major cardiology guidelines, every patient with a diagnosis of nonvalvular atrial fibrillation and risk factors contributing to an increased risk of stroke should be treated with a DOAC.[2,3] However, at least one third of eligible patients are not on a DOAC and uptake trends have been lower than expected.[4,5] Risk stratification tools are essential for guiding anticoagulation therapy to prevent stroke. The CHA_2DS_2-VASc risk score is a validated and guideline recommended score for estimating stroke risks in AF patients.[2,3,6]

Accurate calculation of the CHA_2DS_2-VASc risk scores requires comprehensive assessment of multiple clinical risk factors, necessitating a thorough evaluation of both structured electronic health record data and unstructured clinical notes. While structured electronic health record (EHR) data elements, such as diagnostic billing codes and laboratory results, capture some of these factors, many pertinent details are documented only in unstructured clinical notes.[7] Current practices often rely on manual extraction of information from unstructured notes or approximation of risk scores based solely on structured data. This approach is time-consuming and may lead to incomplete or inaccurate risk stratification, potentially impacting clinical decision-making and patient outcomes. Moreover, the inability to efficiently extract detailed risk factor information from unstructured notes poses a significant barrier to large-scale population health management and research efforts.

Advancements in natural language processing (NLP) offer promising solutions to these challenges. Large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text, enabling more sophisticated extraction of clinical information from unstructured data sources. Retrieval-Augmented Generation (RAG) architectures combine information retrieval with LLMs to enhance the extraction of specific information from unstructured text without the need to fine-tune the LLM. However, despite these technological advancements, the application of RAG and LLMs in extracting specific clinical factors for risk stratification in AF patients remains underexplored. There is a pressing

need for validated methods that harness the rich information contained in clinical notes to improve risk assessment and guide therapy.

To address this gap, we developed and validated a RAG approach paired with the Llama3.1 LLM to extract CHA_2DS_2-VASc risk factors from unstructured clinical notes in patients with AF. We hypothesize that our RAG approach will enhance stroke risk assessment by incorporating detailed clinical information documented in unstructured notes, thereby facilitating more accurate and comprehensive risk stratification.

Methods

Study Design and Setting

We conducted a retrospective study using data from the Yale New Haven Health System (YNHHS) EHR from 2013 to 2024 and the Medical Information Mart for Intensive Care IV (MIMIC-IV) database from 2008 to 2022 for external validation. The Yale Institutional Review Board approved the study protocol and waived the need for informed consent due to the retrospective nature of the study and the use of de-identified data. Access to the MIMIC-IV database was granted following completion of the required training and data use agreements. The Yale Institutional Review Board reviewed the study, approved the protocol, and waived the need for informed consent as this is a secondary analysis of existing data.

Data Sources

We extracted both structured and unstructured data from the YNHHS EHR system and the MIMIC-IV database. The YNHHS system encompasses a large academic hospital and affiliated outpatient clinics serving a diverse patient population representative of national demographics. The MIMIC-IV database is a large, freely available database comprising de-identified health-related data associated with over 200,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2008 and 2022.[8] Structured data included patient demographics, diagnoses, procedures, laboratory results, and medications. Unstructured data comprised clinical notes such as history and physical examinations, progress notes, discharge summaries, and consultation notes, providing detailed narrative documentation of patient encounters.

Study Population

We identified a cohort of patients with AF who had at least one healthcare encounter within YNHHS or documented in the MIMIC-IV database during the study period. Inclusion criteria required patients to be 18 years of age or older and to have a diagnosis of atrial fibrillation, identified using the International Classification of Diseases, Ninth and Tenth Revision (ICD-9 code 427.31 and ICD-10 codes I48.x). Exclusion criteria included patients with incomplete records or those without available unstructured clinical notes. For external validation, we utilized

the MIMIC-IV database, randomly selecting 100 clinical notes from patients with any prior documented diagnosis with AF.

Study Outcomes

The primary outcome of the study was to evaluate the performance of the RAG approach, paired with the Llama3.1 LLM, in extracting individual risk factors relevant to the CHA_2DS_2-VASc risk score from unstructured clinical notes. Secondary outcomes included assessing the accuracy of the calculated CHA_2DS_2-VASc scores when incorporating risk factors identified by the RAG approach compared to scores calculated using structured data alone and evaluating the potential reclassification of patients into different risk categories based on the inclusion of unstructured data.

Model Development

We implemented a RAG approach to extract risk factors relevant to the CHA_2DS_2-VASc scores from unstructured clinical notes without the need to train new models. Our method integrated pre-existing natural language processing models and tools to process the clinical text data effectively. We utilized the pre-trained embeddings model from Alibaba's Institute for Intelligent Computing, GTE (General Text Embeddings) version 1.5.[9] These embeddings capture the semantic meaning of the text and are essential for efficient retrieval. The embeddings were stored and managed using Chroma, a vector database optimized for similarity search.

For each component of the CHA_2DS_2-VASc risk score, we formulated specific queries to retrieve relevant information from the clinical notes. Using cosine similarity, we retrieved the top 10 most relevant text chunks for each query from the vector database. To enhance the relevance of the retrieved text chunks, we employed a pre-trained CrossEncoder model from the Beijing Academy of Artificial Intelligence as a reranker.[10] This model evaluated the relevance of each retrieved text chunk concerning the specific query. Subsequently, we used the Llama3.1 LLM to generate answers to the predefined queries based on the reranked text chunks.[11] The final output for each clinical note was a set of binary “Yes” or “No” answers corresponding to each component of the CHA_2DS_2-VASc risk score. By aggregating these results, we efficiently extracted the necessary risk factors from unstructured clinical notes.

Manual Annotation and Validation

To establish a gold standard for model evaluation, two clinicians independently reviewed and manually annotated CHA_2DS_2-VASc risk factors in random subsets of clinical notes from both datasets. Specifically, they annotated 200 clinical notes from the YNHHS cohort and 100 clinical notes from the MIMIC-IV database. Each risk factor component was labeled according to predefined criteria, such as the presence of hypertension or history of stroke. Performance metrics were calculated separately for both datasets to assess the model's generalizability. Furthermore, we deployed our method on the remaining 800 unannotated clinical notes from YNHHS to extract risk factors and calculate CHA_2DS_2-VASc scores. This allowed us to evaluate the model's performance in a real-world setting and assess the potential impact on patient risk stratification in a healthcare system. Performance metrics calculated for each risk factor included sensitivity (recall), specificity, positive predictive value (precision), F1-score, accuracy, and area under the receiver operating characteristic curve (AUROC). These metrics provided a comprehensive assessment of the RAG approach's accuracy in extracting risk factors from unstructured text.

Calculation of Risk Scores

For each patient in both the YNHHS and MIMIC-IV datasets, we calculated the CHA_2DS_2-VASc scores using two methods. The first method utilized structured data alone, identifying risk factors from structured EHR diagnostic codes. The second method incorporated additional risk factors extracted from unstructured clinical notes by the RAG model. By comparing the scores obtained from both methods, we assessed the impact of including unstructured data on overall risk stratification. Patients were subsequently categorized into risk groups based on established thresholds for each scoring system.

Statistical Analysis

Descriptive statistics were used to summarize patient demographics and clinical characteristics in both the YNHHS and MIMIC-IV datasets. Continuous variables were expressed as means with standard deviations, while categorical variables were reported as frequencies and percentages. To assess the performance of the RAG model in extracting risk factors from unstructured clinical notes, we calculated several metrics for each CHA_2DS_2-VASc component. These metrics

included sensitivity (recall), specificity, positive predictive value (precision), F1-score, accuracy, and area under the receiver operating characteristic curve (AUROC). These calculations were performed separately for the annotated subsets of both YNHHS (200 notes) and MIMIC-IV (100 notes) datasets to evaluate the model's performance and generalizability. All statistical analyses and data processing were performed using Python version 3.10, utilizing appropriate libraries for natural language processing, machine learning, and statistical computations.

Results

Study Population

A total of 1,000 clinical notes from 934 unique patients at YNHHS were used in the analysis. The mean age of the patients was 66.9 years (SD ± 14.8). In this cohort, 446 (44.6%) individuals were female. Regarding race and ethnicity, 778 (77.8%) individuals were White or Caucasian, 152 (15.2%) were Black or African American, 5 (0.5%) were Asian, and 65 (6.5%) were classified as Other or Unknown. Among the comorbidities analyzed, hypertension was the most prevalent, affecting 764 (76.5%) patients. Diabetes mellitus was present in 403 (40.3%) patients, while congestive heart failure was observed in 549 (55.0%) patients. Additionally, 428 (42.8%) patients had a history of stroke, transient ischemic attack (TIA), or thromboembolism, and 522 (52.3%) patients had vascular disease (**Table 1**).

Performance of the RAG Model in Extracting Risk Factors

The RAG model demonstrated robust performance in extracting individual risk factors from unstructured clinical notes. Analysis of the full set of 1,000 clinical notes revealed that the RAG model consistently identified several risk factors more frequently than extraction from structured data alone. Hypertension was detected in 82.4% of notes using the RAG model, compared to 26.2% using structured data. Similarly, the RAG model identified a history of stroke or transient ischemic attack in 62.9% of notes, versus 45.5% in structured data. Vascular disease was recognized in 83.4% of cases by the RAG model, while structured data identified it in only 56.6% of cases. The model also outperformed structured data in identifying diabetes mellitus (84.1% vs. 47.2%) (**Table 2**).

Validation Against Expert-Annotated Notes

In the subset of 200 expert-annotated clinical notes from YNHHS, hypertension was identified with a sensitivity of 99% and specificity of 93%, resulting in an F1-score of 99% and an AUROC of 96%. For diabetes mellitus, the model achieved a sensitivity of 98%, specificity of 97%, F1-score of 98%, and AUROC of 97%. In identifying age-related risk factors, age ≥75 years was detected with 100% sensitivity and 96% specificity (F1-score: 98%, AUROC: 98%) (**Table 3**). Similar results were seen in the 100 annotated clinical notes from MIMIC IV, with AUROCs ranging from 0.95 to 0.99 for these key risk factors.

Discussion

In this study, we developed and validated a RAG model paired with the Llama3.1 LLM to extract CHA_2DS_2-VASc risk factors from unstructured clinical notes in patients with AF. Our findings demonstrate that the RAG model significantly enhances the identification of critical risk factors compared to extraction from structured data alone. Specifically, the model identified hypertension in 82.4% of notes versus 26.2% from structured data, and similar improvements were observed for other risk factors such as stroke/TIA and vascular disease. The model exhibited high sensitivity and specificity when validated against expert-annotated notes, with AUROC values exceeding 96% for key risk factors. Incorporating the RAG-extracted risk factors led to changes in calculated risk scores and reclassification of patients into different risk categories, highlighting the model's potential to improve risk stratification in clinical practice.

Our study builds upon previous efforts to leverage NLP for extracting clinical information from EHRs. Traditional methods, such as rule-based algorithms and basic machine learning models, often struggle with the complexity and variability of unstructured clinical text.[12] While some studies have applied NLP techniques to extract specific clinical entities, they frequently rely on extensive manual feature engineering and may not generalize well across different datasets. In contrast, our RAG approach combines information retrieval with advanced language modeling, enabling it to capture nuanced clinical information embedded in free-text notes without the need for extensive manual curation. This represents a significant advancement in the automated extraction of clinically relevant data for risk assessment in AF patients.

The implications of our findings are substantial for both clinical practice and research. By enhancing the accuracy of risk factor identification, our model can improve the calculation of CHA_2DS_2-VASc scores, leading to more precise risk stratification and potentially better-informed decisions regarding anticoagulation therapy. This is particularly important given that underestimation of risk can result in inadequate prophylaxis against stroke, while overestimation may expose patients to unnecessary bleeding risks. Additionally, the ability to automatically extract detailed clinical information from unstructured notes can streamline workflow efficiencies, reduce the burden on clinicians, and facilitate large-scale population health management and research initiatives. Integrating such models into EHR systems could also support clinical decision support tools, providing clinicians with accurate risk assessments at the point of care.

Despite the promising results, our study has limitations that warrant consideration. While the model demonstrated high performance for certain risk factors, the extraction of other components and other risk scoring measures, such as HAS-BLED, was not fully reported and may require further optimization. Additionally, the impact of incorporating unstructured data on actual clinical outcomes was not assessed in this study. Prospective studies are needed to evaluate whether improved risk stratification translates into better patient outcomes.

Conclusion

An LLM-optimized RAG can accurately extract CHADS-VASc risk factors from unstructured clinical notes in AF patients. This approach can enable computable risk assessment and guide appropriate anticoagulation therapy

Data Sharing Statement

The data used for this study cannot be publicly shared as it represents protected health information and sharing data will be a violation of patient privacy. The MIMIC-IV cohort has an application for access at https://physionet.org/content/mimiciv/3.0/.

Funding

PA is supported by F30HL176149. RK is supported by the National Heart Lung and Blood Institute of the National Institutes of Health (under awards R01HL167858 and K23HL153775) and the Doris Duke Charitable Foundation (under award 2022060). Dr. Thangaraj was supported by the National Institutes of Health under award T32HL155000. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclosures

Dr. Khera is an Associate Editor of JAMA. He also receives research support, through Yale, from Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. He is a coinventor of U.S. Pending Patent Applications 63/562,335, 63/177,117, 63/428,569, 63/346,610, 63/484,426, 63/508,315, and 63/606,203. He is a co-founder of Ensight-AI, Inc. and Evidence2Health, health platforms to

improve cardiovascular diagnosis and evidence-based cardiovascular care. Dr. Krumholz reported receiving expenses and/or personal fees from UnitedHealthcare, Element Science, Inc, Aetna Inc, Reality Labs, Tesseract/4 Catalyst, F-Prime, the Siegfried & Jensen law firm, the Arnold & Porter law firm, and the Martin Baughman, PLLC; being a cofounder of Refactor Health and Hugo Health; and being associated with contracts through Yale New Haven Hospital from the Centers for Medicare & Medicaid Services and through Yale University from Johnson & Johnson outside the submitted work. Dr. Thangaraj is a coinventor of a provisional patent (63/606,203). The other authors report no relevant disclosures.

Figure 1: Study Flow Diagram

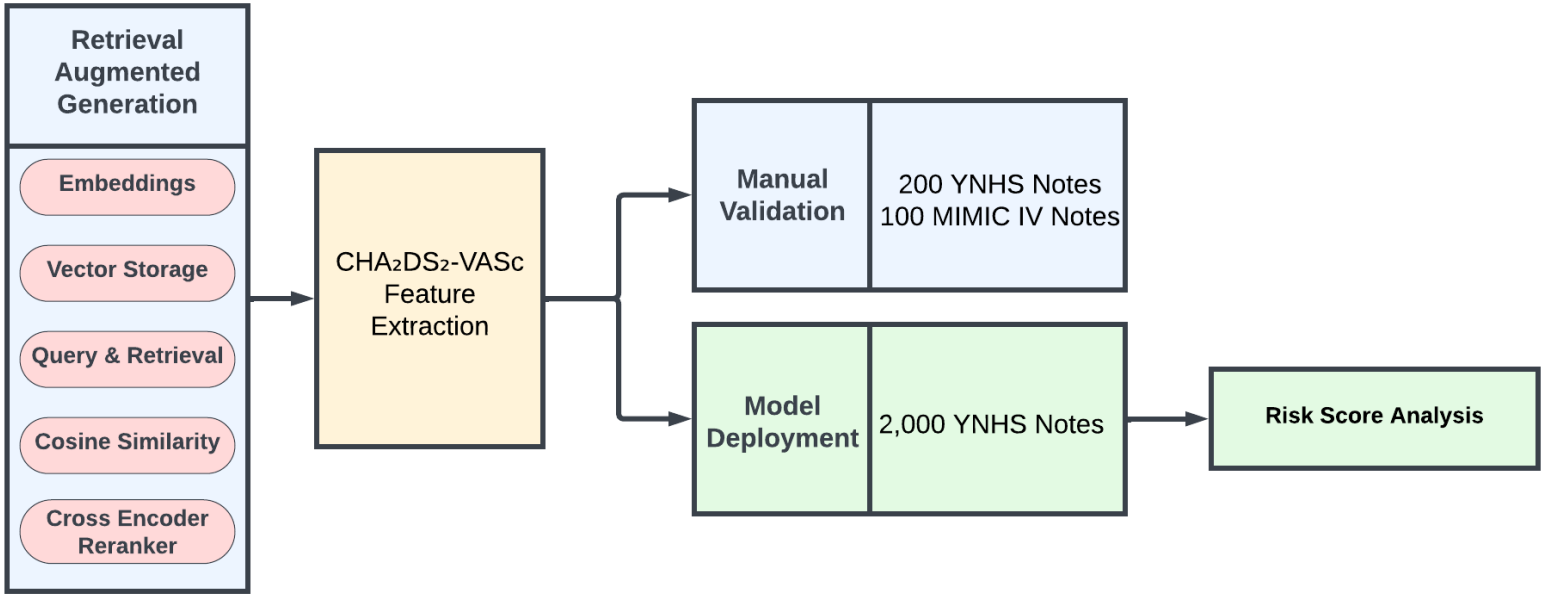

Figure 2: Comparison of Performance Metrics CHA_2DS_2-VASc Feature Extraction. A grouped bar chart comparing the precision, recall, F1-score, and accuracy for each risk factor relevant to the CHA_2DS_2-VASc.

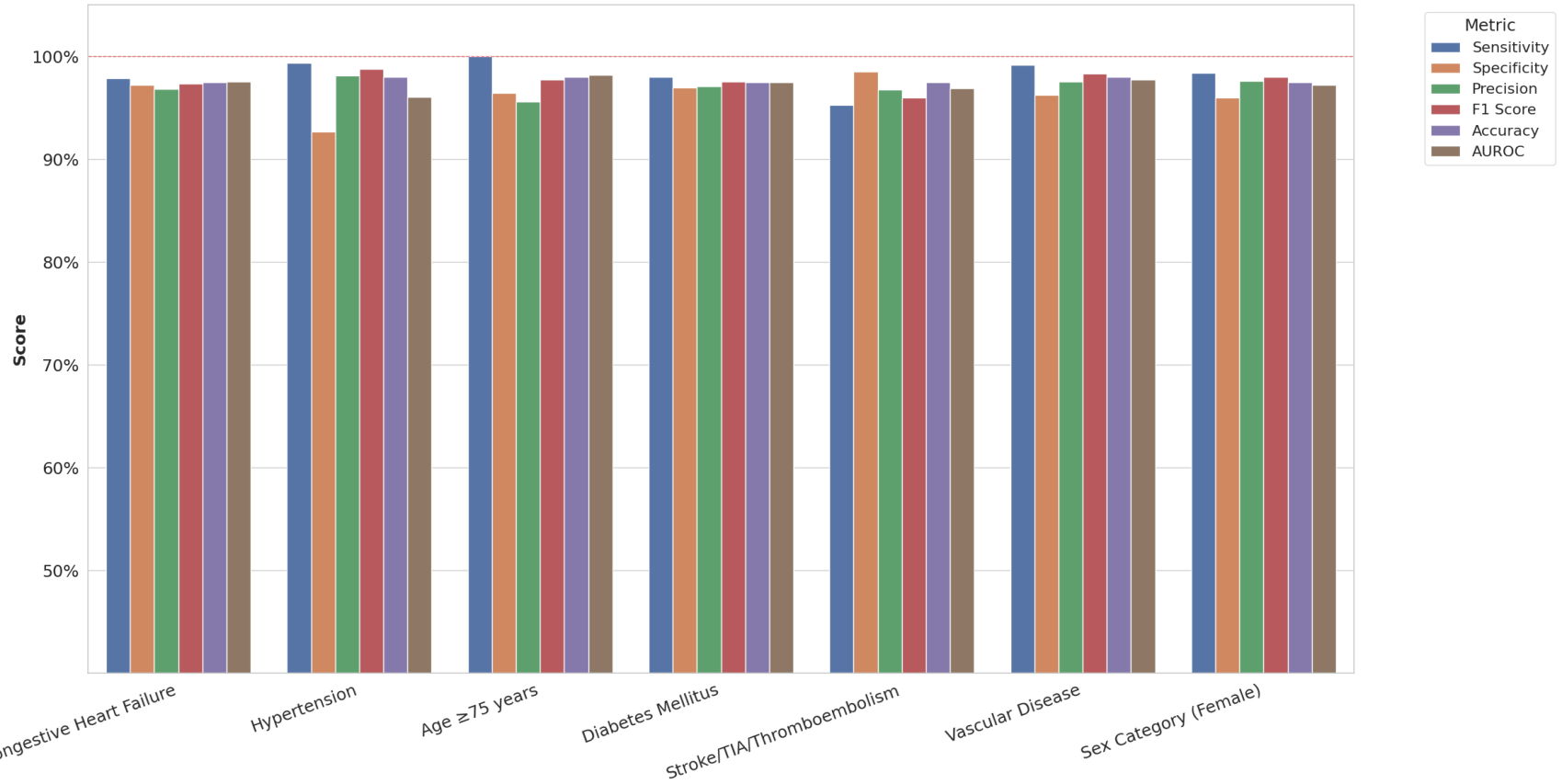

Table 1: Baseline Characteristics of the Study Population

Characteristic	**Yale New Haven Hospital (n = 250 patients, 1,000 notes)**
Age, mean ± SD (years)	64.9 (± 14.0)
Female sex (%)	505 (50.5%)
Race/Ethnicity	
White or Caucasian (%)	675 (67.5%)
Black or African American (%)	241 (24.1%)
Asian (%)	2 (0.2%)
Other/Unknown (%)	82 (8.2%)
Comorbidities	
Congestive Heart Failure (%)	695 (69.6%)
Hypertension (%)	824 (82.4%)
Diabetes Mellitus (%)	471 (47.1%)
Stroke/TIA/Thromboembolism (%)	454 (45.4%)
Vascular Disease (%)	565 (56.5%)

Table 2: Comparison of Risk Factor Identification Between RAG Model and Structured Data

Risk Factor	RAG Model Identified	Structured Data Identified	Cohen's Kappa*
Congestive Heart Failure	793 (79.3%)	696 (69.6%)	0.28
Hypertension	262 (26.2%)	824 (82.4%)	0.06
Diabetes Mellitus	841 (84.1%)	472 (47.2%)	0.13
Stroke/TIA/Thromboembolism	629 (62.9%)	455 (45.5%)	0.21
Vascular disease	834 (83.4%)	566 (56.6%)	0.15

Table 3: Performance Metrics of the RAG Model in Extracting Risk Factors

Risk Factor	Sensitivity	Specificity	Precision	F1-Score	AUROC
Congestive Heart Failure	0.98	0.97	0.97	0.97	0.98
Hypertension	0.99	0.93	0.98	0.99	0.96
Age ≥75 years	1.00	0.96	0.96	0.98	0.98
Diabetes Mellitus	0.98	0.97	0.97	0.98	0.97
Stroke/TIA/Thromboembolism	0.95	0.99	0.97	0.96	0.97
Vascular disease	0.99	0.96	0.98	0.98	0.98
Sex Category (Female)	0.98	0.96	0.98	0.98	0.97

References

1. Alshehri, A. M. Stroke in atrial fibrillation: Review of risk stratification and preventive therapy. *J. Family Community Med.* **26**, 92–97 (2019).
2. Van Gelder, I. C. *et al.* 2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). *Eur. Heart J.* (2024) doi:10.1093/eurheartj/ehae176.
3. Joglar, J. A. *et al.* 2023 ACC/AHA/ACCP/HRS guideline for the Diagnosis and Management of Atrial Fibrillation: A report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. *Circulation* **149**, e1–e156 (2024).
4. Navar, A. M. *et al.* Trends in oral anticoagulant use among 436 864 patients with atrial fibrillation in community practice, 2011 to 2020. *J. Am. Heart Assoc.* **11**, e026723 (2022).
5. Wheelock, K. M. *et al.* Clinician trends in prescribing direct oral anticoagulants for US Medicare beneficiaries. *JAMA Netw. Open* **4**, e2137288 (2021).
6. Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A. & Crijns, H. J. G. M. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* **137**, 263–272 (2010).
7. Adejumo, P. *et al.* A Deep Learning Approach for Automated Extraction of Functional Status and New York Heart Association Class for Heart Failure Patients During Clinical Encounters. *medRxiv* (2024) doi:10.1101/2024.03.30.24305095.
8. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

9. Zhang, X. *et al.* MGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv [cs.CL]* (2024).
10. Chen, J. *et al.* BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv [cs.CL]* (2024).
11. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv [cs.CL]* (2023).
12. Pandey, B., Kumar Pandey, D., Pratap Mishra, B. & Rhmann, W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *J. King Saud Univ. - Comput. Inf. Sci.* **34**, 5083–5099 (2022).