Probabilistic Inference of Comorbidities from Symptoms in Patients with Atrial Fibrillation: An Ontology-Driven Hybrid Clinical Decision Support System

Alexander Lacki[1], Diego Bosca[2], Antonio Martinez-Millana[1]

[1] ITACA Institute, Universitat Politecnica de Valencia, Valencia, Spain
[2] Veratech for Health S.L., Valencia, Spain

Abstract

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia. While AF is a cardiological disease, its risk factors and mechanisms are often rooted in non-cardiological comorbidities, introducing complexity in the treatment of the heterogeneous patient population.

This study presents the development of a clinical decision support system (CDSS), which aims to mitigate potential challenges of the cross-disciplinarity of AF. A knowledge base is implemented to capture the hierarchical nature of relevant concepts. Naïve Bayes classifiers are used to predict the patient comorbidities related to AF mechanisms and risk factors based on provided symptoms. The resulting CDSS infers comorbidities with a top-k accuracy of 0.53, 0.80, and 0.88 for k = 1, 3, and 5 respectively.

1. Introduction

AF is the most prevalent cardiac arrhythmia affecting more than 33 million patients globally [1]. It is associated with advanced age, an abundance of comorbidities, such as systemic diseases, endocrine disorders, metabolic disturbances, and genetic conditions, creating a pathophysiological foundation that goes beyond the scope of cardiology [2]. As a result, current guidelines recommend an integrated treatment approach combining the expertise of cross-disciplinary teams consisting of a variety of specialists [3].

Probabilistic methods to infer patients' diseases based on Bayes' theorem have been employed since the earliest days of biomedical informatics and have shown promising results in many studies [4]. In recent years, such data-driven methods have increasingly been complemented with clinical knowledge in the form of domain ontologies, introducing semantics in the form of relationships between concepts and allowing for semantic reasoning and improving operations on unstructured data [4].

The objective of this study is the development of a CDSS that predicts AF related diagnoses from reported symptoms while considering prediction uncertainty, and making use of domain knowledge to account for structured and unstructured data.

2. Methods

2.1. Data and Cohort Definition

The MIMIC-III database v1.4 [5], a single center critical care database is used. It contains deidentified clinical data from 53,423 ICU admission of 38,597 individual adult patients collected from critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. Patients with a diagnostic code indicating the presence of AF (ICD-9: 427.31) are included in the analysis. For patients with more than one registered ICU visit, only the first visit with the accompanying diagnostic code is considered.

2.2. Knowledge Base

To facilitate the semantic reasoning on symptoms and diagnoses, which are inherently of a hierarchical nature, a knowledge base is developed. Protege [6], an ontology editor, is used to incorporate relevant concepts into their corresponding hierarchies. Concepts are extracted from SNOMED CT and annotated with object properties such as the corresponding Unified Medical Language System (UMLS) [7] identifiers, and ICD-9 codes. Further, an object property is implemented indicating the relevance of concepts within the hierarchy. This object property allows for a reduction of groups of concepts to a superclass that is considered to be of importance.

2.3. Data Extraction and Natural Language Processing

Symptoms are identified from free-text clinical notes using a SciSpacy's "en_core_sci_sm" NLP pipeline [8]. The "en_core_sci_sm" pipeline performs token vectorization,

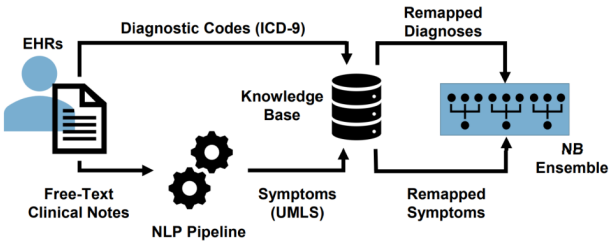

Figure 1. Workflow schematic.

part-of-speech tagging, parsing, attribute ruling, lemmatization, and named entity recognition. The named entities are linked to UMLS identifiers using the corresponding entity linker. Diagnostic codes are captured from the corresponding database tables. The identified symptoms, in the form of UMLS identifiers, and diagnoses, in the form of ICD codes, are mapped to the knowledge base and the relevant superclasses within the hierarchy are inferred.

2.4. Probabilistic Inference

The obtained dataset of symptoms and corresponding diagnoses for each patient is employed to train an ensemble of naïve Bayes classifiers using Scikit-learn [9]. Each classifier in the ensemble predicts the probability of a given diagnosis using the present symptoms. Predictive performance is assessed in terms of the following metrics using 5-fold cross validation:

- **Top-k Accuracy**: The probability of a present diagnosis being among the k predictions (ranked by the predicted probability).
- **Top-k Precision**: The proportion of predicted diagnoses being present within the top k diagnoses (ranked by the predicted probability).

Figure 1 presents a schematic of the entire workflow.

3. Results

10,277 patients meeting the inclusion criteria were identified in the database. Patients had 7.58±4.19 symptoms, and 2.92±1.96 diagnoses. A total of 32 symptom groups and 18 diagnostic groups were incorporated into the knowledge base. The prevalence of the captured diagnoses and symptoms are presented in figures 2 and 3, and histograms are shown in figure 4.

Hypertension, heart failure, and atherosclerosis were the most commonly encountered diagnostic groups within the cohort, being present in 48.2%, 43.3%, and 36.5% of patients respectively. Beyond cardiovascular comorbidities, patients commonly suffered from renal insufficiency (34.5%), diabetes (22.9%), anemia (11.8%), sepsis (11.7%), and thyroid disorders (11.0%).

The majority of patients exhibited symptoms within the categories of swelling, dynspnea, and cough, with 96.3%, 80.8% and 69.6% of patients having symptoms within the groups, respectively.

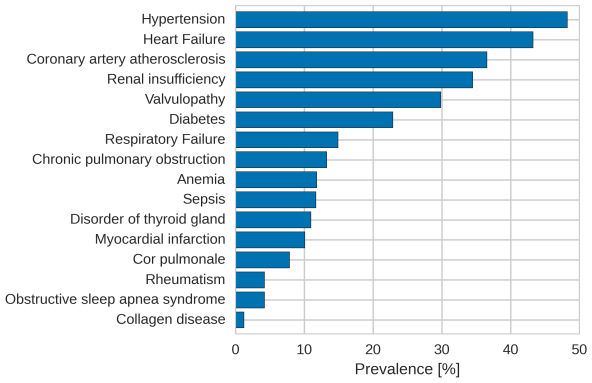

Figure 2. Prevalence of diagnosis groups within the patient cohort.

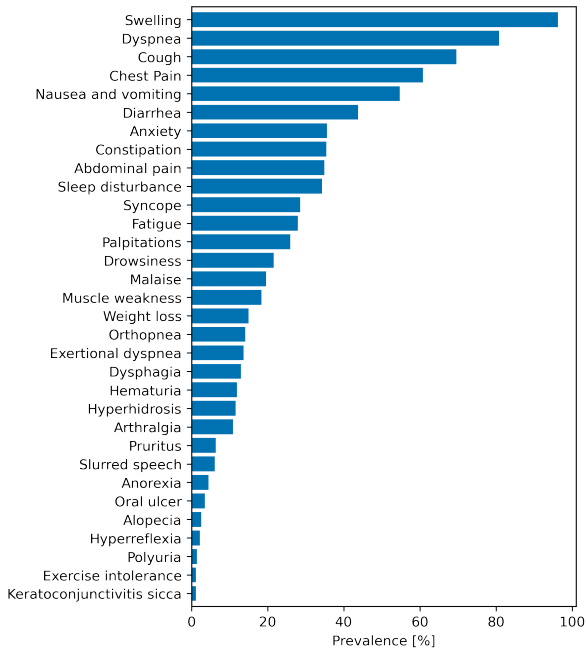

Figure 3. Prevalence of symptom groups within the patient cohort.

Cross validation of the ensemble of naïve Bayes classifiers demonstrated a a top-k accuracy of 0.53, 0.80, and 0.88, and a top-k precision of 0.53, 0.44, and 0.37, for k = 1, 3, and 5 respectively. The two metrics are visualized in figure 5.

4. Discussion

The aim of this study was the development of a decision support system that combines predictive performance, while accounting for prediction uncertainty and accounting for domain knowledge.

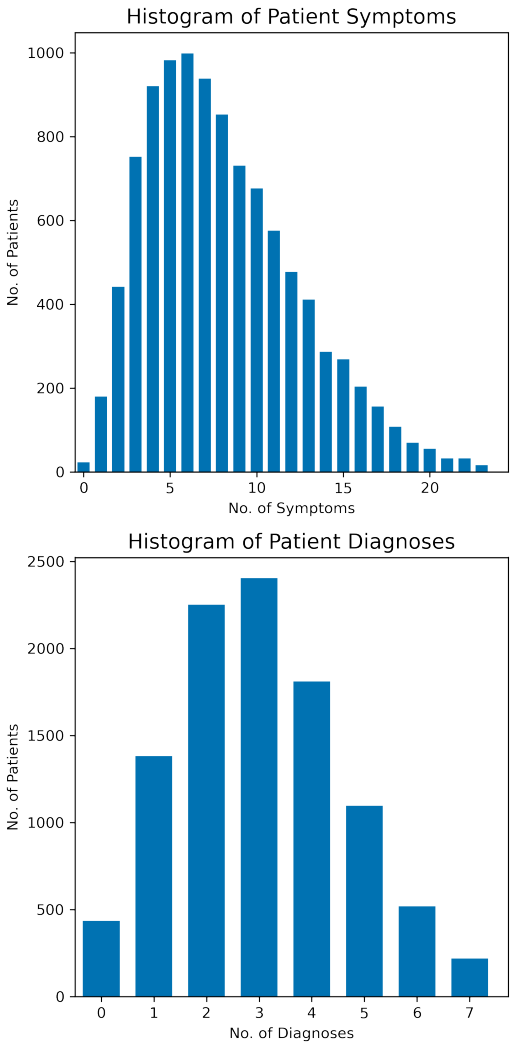

Figure 4. Histograms of patient symptoms (top) and diagnoses (bottom).

The proposed approach successfully integrates structured and unstructured data, to capture concepts not readily available in tabular form. The use of an NLP pipeline allows for the identification of relevant entities from free-text medical notes, and may be extended to identify further concepts such as signs, observations, and measurements.

The use of a domain ontology allows for the representation of the inherently hierarchical nature of symptoms and diagnoses. Further, it provides a basis that can readily be augmented with additional variables and functionalities. A natural extension would be the inclusion of possible diagnostic tests, which, when annotated with the corresponding price, could allow for a recommendation of the most cost effective diagnostic sequence.

The resulting CDSS provides predictions while accounting for uncertainty, and shows encouraging predictive performance. It should, however, be noted that the performance was evaluated on the development dataset, and performance in an external cohort may be inferior.

A possible limitation of the presented method is the choice of the naïve Bayes classifier. Naïve Bayes classifiers make the assumption of feature independence, which is unlikely to hold true within the scope of this work. Even though this impediment may be assumed to degrade the predictive performance of the employed models, previous studies have shown the naïve Bayes classifier to be robust and to show similar performance to models accounting for conditional dependence [10].

The results obtained in this study provide encouragement for further studies exploring the applicability of the presented framework. While this work used an ICU cohort to develop a proof of concept, future work should evaluate community cohorts, which are arguably a more appropriate use case. Further variables need to be included that could extend beyond symptoms and diagnoses, and include observations and measurements, which could further streamline the diagnostic process.

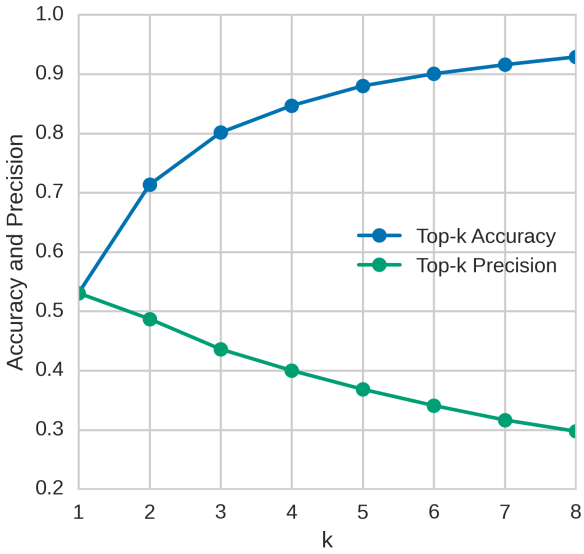

Figure 5. Top-k accuracy (blue) and top-k precision (green) of the naïve Bayes ensemble.

5. Conclusions

The present work demonstrates the ability of probabilistic classifiers to identify AF related comorbidities from reported symptoms in clinical notes. A CDSS such as the one presented could aid clinicians in the identification of comorbidities and mechanisms driving the arrhythmia, and provide support in the selection of the optimal diagnostic and treatment strategies, reducing the strain on specialists, while streamlining diagnostic processes.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 860974. This publication reflects only the authors' view, and the funding Agencies are not responsible for any use that maybe made of the information it contains.

References

[1] Rahman F, Kwan GF, Benjamin EJ. Global epidemiology of atrial fibrillation. Nature Reviews Cardiology 11 2014; 11:639–654.

[2] Wijesurendra RS, Casadei B. Mechanisms of atrial fibrillation. Heart 12 2019;105:1860–1867.

[3] Hindricks G, et al. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European association for cardio-thoracic surgery (EACTS). European Heart Journal 2 2021;42:373–498.

[4] Musen MA, Middleton B, Greenes RA. Biomedical Informatics. Springer International Publishing, 2021; 795–840.

[5] Johnson AE, et al. MIMIC-III, a freely accessible critical care database. Scientific Data 12 2016;3:160035.

[6] Musen MA. The protégé project: a look back and a look forward. AI Matters 2015;1(4):4–12.

[7] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Research 1 2004;32:267–270.

[8] Neumann M, et al. Scispacy: Fast and robust models for biomedical natural language processing. Association for Computational Linguistics, 2019; 319–327.

[9] Pedregosa F, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011; 12:2825–2830.

[10] Eisenstein EL, Alemi F. An evaluation of factors influencing Bayesian learning systems. Journal of the American Medical Informatics Association 5 1994;1:272–284.

Address for correspondence:

Alexander Lacki
ITACA Institute. Universitat Politécnica de València.
Edificio 8G acceso B. Camino de Vera s/n.
46022 Valencia, Spain.
alacki@upvnet.upv.es