



Published in final edited form as:

*Health Informatics J.* 2024 ; 30(4): 14604582241300025. doi:10.1177/14604582241300025.

## Integrating data-driven and knowledge-driven approaches to analyze clinical notes with structured data for sarcopenia detection

Xiao Luo, Ph.D.<sup>1,2,\*</sup>, Haoran Ding, M.A.<sup>3,\*</sup>, Stuart J. Warden, Ph.D.<sup>4,5</sup>, Ranjani N. Moorthi, M.D.<sup>5,6</sup>, Erik A. Imel, M.D.<sup>5,6</sup>

<sup>1</sup>Department of Management Science and Information Systems, Oklahoma State University, Oklahoma, USA

<sup>2</sup>Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, Indiana, USA

<sup>3</sup>Department of Electrical and Computer Engineering, Purdue University Indianapolis, Indianapolis, Indiana, USA

<sup>4</sup>Department of Physical Therapy, Indiana University School of Health and Human Sciences, Indianapolis, Indiana, USA

<sup>5</sup>Indiana Center for Musculoskeletal Health, Indiana University School of Medicine, Indianapolis, Indiana, USA

<sup>6</sup>Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA

### Abstract

**Background:** Patients with sarcopenia often go undetected in busy clinical practices since the muscle measurements are not easily incorporated into routine clinical practice. The current research fills the gap by utilizing unstructured clinical notes combined with structured data from electronic health records (EHR), to increase sarcopenia detection.

**Methods:** We developed and evaluated four approaches to first extract clinical note features, then integrate with structured data for sarcopenia detection models. Case studies were used to demonstrate the interpretation of the results and show the important association between predictors and outcomes.

---

**Corresponding Author address:** Xiao Luo, Department of Management Science and Information Systems, Oklahoma State University, Oklahoma, USA, xiao.luo@okstate.edu.

\*Xiao Luo and Haoran Ding are Co-first authors.

**Contributorship:** EAI and XL had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. The study was conceptualized and designed by XL, SJW, RNM, and EAI; funding was obtained by XL, RNM, and EAI. Data collection and cleaning was done by HD, SJW, RNM, EAI, and XL. Analysis was by XL, HD, RNM, and EAI. Interpretation was by XL, RNM, SJW, and EAI. The first draft was completed by XL, RNM, and EAI, and subsequent versions edited and approved by all authors.

**Conflicts of interest and financial disclosures:** All authors declare that they have no conflicts of interest.

**Ethical Statement:** This study received ethical approval from the Indiana University IRB (approval #2004191295) on 04/16/2020. This is an IRB-approved retrospective study, all patient information was de-identified and patient consent was not required. Patient data will not be shared with third parties.

**Results:** Out of 1304 participants, 1055 were controls, 249 met at least one criterion for Sarcopenia. The best performing model which incorporated both data-driven and knowledge-driven approaches to integrate clinical note features demonstrated a higher mean area under the curve (AUC = 73.93%, (95%CI, 73.83–74.02)) compared to the baseline model (AUC 71.59%, (95%CI, 71.56–71.61)). The case study shows that the important clinical note predictors may contribute to detection of sarcopenia such as “cane”, “walker”, “unsteady”, etc.

**Conclusions:** Incorporating clinical note features in sarcopenia detection models can identify a greater number of patients at risk for sarcopenia, potentially leading to targeted muscle testing assessments and corresponding treatments to address sarcopenia.

## Keywords

electronic health records; feature selection; natural language processing; predictive modeling; sarcopenia

---

## 1. Introduction

Sarcopenia is a condition of generalized low muscle mass and strength and poor physical performance that increases with aging and complicates many chronic diseases<sup>1,2</sup> including renal diseases, cancers, etc. The etiology of sarcopenia is multifactorial, including effects of age, chronic disease, inflammation, neurologic disorders, physical inactivity, malnutrition, and obesity. Sarcopenia increases the risk for multiple costly health outcomes including falls, fractures, mobility disorders, impairments in activities of daily living, loss of independence, and long-term care placement<sup>3–10</sup>. Research also shows that the resulting restrictions in activities of daily living and independence could have downstream effects on social engagement and psychological status<sup>11,12</sup>. The age-adjusted mortality rate is also >2.3 times higher among sarcopenic compared to non-sarcopenic older adults<sup>7,13–15</sup>. Sarcopenia also increases the risk for hospitalization and results in 2-fold increased costs for hospitalizations and for long-term care due to disability<sup>10,16,17</sup>.

To promote early clinical detection and treatment of patients having or at risk for sarcopenia, the European Working Group on Sarcopenia in Older People (EWGSOP2) presented consensus definitions for sarcopenia, characterizing probable sarcopenia by low muscle strength, confirmed sarcopenia by the presence of low muscle mass, and severe sarcopenia as the combination of low muscle strength, mass, and physical performance parameters<sup>2</sup>. However, patients with sarcopenia often go undetected in busy clinical practices<sup>18</sup>. First, sarcopenia diagnosis requires measuring muscle mass, strength, and performance<sup>1,2,18</sup>, necessitating access to specialized equipment to measure muscle mass, such as magnetic resonance imaging (MRI), computed tomography (CT), or whole-body dual-energy x-ray absorptiometry (DXA), or dynamometry to measure grip strength. In addition, some tests such as gait speed require available space and time for assessment. The muscle mass measurement exams are not easily incorporated into routine office clinical practice, where primary providers have time-limited clinical encounters, and sometimes limited knowledge of sarcopenia and of its assessment through objective measures (e.g., grip strength)<sup>18</sup>. Sarcopenia did not have an ICD code until 2016<sup>19</sup>. The previous research<sup>20</sup> combines natural language processing (NLP) techniques based on key words and phrases with expert

review of text terms from clinical notes supporting sarcopenia (as well as cachexia and frailty, which can be related to sarcopenia) along with ICD codes, to identify patients with sarcopenia in the EHR. While this mechanism enabled some identification of sarcopenia, this mechanism showed low performance in detecting sarcopenia. When relying on those few text terms and ICD codes, only those patients having these in their notes are detectable. However, sarcopenia is under-diagnosed. In another study we found less than 2% of those meeting sarcopenia criteria in the research testing core had the ICD10 code or text term in their clinical record<sup>21</sup>. On the other hand, the occurrence of ICD codes or terms in EHR clinical notes does not guarantee that the condition is present, and objective measurements were not typically documented within the EHR. Therefore, it is important to generate a tool to identify patients with, or at risk for, sarcopenia from data already available within electronic health records (EHR), even when objective measures such as grip strength and gait speed are not recorded during routine clinical practice. Large EHR datasets combining clinical text notes with diagnosis codes and other coded data provide an opportunity to identify patients with sarcopenia on a population level.

We previously used structured data from the EHR to generate machine learning models to detect the presence of sarcopenia criteria, using measurements in the research core as a gold standard<sup>21</sup>. To the best of our knowledge, the study cohort dataset used in the previous research<sup>21</sup> (and also in the current manuscript) has the largest patient cohort with direct measurements for sarcopenia combined with robust structured and unstructured data from a multiple health institution EHR data exchange. Clinical text note extraction manually is cumbersome and natural language processing (NLP) uses computer algorithms to select out key words and phrases. Our objective of the current analysis is to investigate an efficient and effective model to combine NLP derived information from clinical text notes with the coded (structured) data from the EHR to improve detection of sarcopenia. The main contributions of this research include: (1) Investigating and comparing various approaches to integrate the clinical text notes for sarcopenia detection; (2) Developing a model to integrate both the data-driven and knowledge-driven approaches to incorporate text data.

## 2. Materials and Methods

### 2.1 Dataset and preprocessing

This retrospective analysis includes EHR data of patients who were categorized as sarcopenic or nonsarcopenic (controls) according to their muscle or physical performance testing in the a state Center for Musculoskeletal Health's Function, Imaging, and Tissue Resource Core (FIT Core) from the previous study<sup>21</sup>. The testing includes grip strength<sup>3,22</sup>, repeat chair stand test<sup>22</sup>, gait speed<sup>23</sup>, muscle strength<sup>24</sup>, Short Physical Performance Battery (SPPB)<sup>4,5</sup>, and appendicular skeletal muscle mass adjusted for height in meters-squared (ASM/m<sup>2</sup>) as measured by dual-energy X-ray absorptiometry (DXA)<sup>1</sup>. Each test has a criteria threshold to identify sarcopenia. Patients were classified using definitions and thresholds for sarcopenia from the EWGSOP2 guidelines<sup>18</sup>, applied as in the previous publication<sup>21</sup> (See also Supplemental methods). For the current analysis, if a patient met one or more of the criteria test thresholds for sarcopenia, the patient was categorized as having sarcopenia, otherwise, the patient was categorized as a control. In this study, we included

1,304 adult participants in the FIT Core protocol who had EHR data in a state Network for Patient Care between January 2016 and August 2020. The additional inclusion criteria were each patient having a minimum of 2 years of EHR data before the FIT Core visit date for muscle or physical performance testing. There are no exclusion criteria. This study was approved by the institutional review board (IRB) of a state University.

The structured data from the EHR includes diagnosis codes (ICD-10), laboratory test results, medications, and BMI values. For diagnoses, we used the third level to the leaf nodes of the ICD-10 code hierarchy, which groups diagnoses into categories<sup>25</sup>. For example, A00 to A09 are grouped as “intestinal infectious diseases”. We grouped various types of tumors or cancers together as the neoplasms group including ICD-10 codes from C00 to D49. Pregnancy or childbirth-relevant codes (O00 to O9A) were excluded. For medication, we used the drug group specified in the national drug code (NDC) directory and excluded “medical devices and supplies”, “diagnostic products”, “nasal agents- systemic topical”, and “dermatological” categories. The anti-infectious medications are grouped into one category, including “Anti-infective Agents – Misc.”, “Antifungals”, “Antivirals”, “Aminoglycosides”, “Antimycobacterial Agents”, “Cephalosporins”, “Fluoroquinolones”, “Penicillin”, “Tetracyclines” and “Macrolides”. Each diagnosis and medication variable were coded as binary representation (1 or 0) indicating whether the diagnosis was made for the patient or medication was prescribed. For laboratory tests, we included all tests having measures by 10% or more of the patients in our study cohort. The number values of the laboratory test results were then standardized to “low”, “normal” or “high” using the laboratory reference ranges. When multiple results of the same laboratory test were found in the EHR, the most recent result was used. If laboratory tests were not conducted or absent, they were considered as “normal,” operating under the assumption that if any given test result was not available it was more likely to be normal than abnormal, and we would not be able to conclude otherwise. The BMI values collected during the muscle or physician performance testing were used and coded to represent “low”, “normal” and “high” using the published standard<sup>21</sup>. Each laboratory test result or BMI value were coded as 0 or 1 or 2 representing “low”, “normal” or “high”.

Clinical notes include all types of clinical reports stored in the EHR. Because of the amount of textual data, we first applied UMLS MetaMap<sup>26</sup> to select the sentences containing concepts of selected semantic categories including ‘Disease or Syndrome’, ‘Diagnostic Procedure’, ‘Activity’, ‘Daily or Recreational Activity’, ‘Body Part, Organ, or Organ Component’, ‘Body Space or Junction’, ‘Medical Device’, and ‘Sign or Symptom’. Figure 1 shows a clinical report snippet with selected sentences containing concepts in some of these semantic categories of the UMLS MetaMap. The selected sentences in the clinical notes were processed to generate word-based n-grams. 1-grams, 2-grams, and 3-grams were generated. Each n-gram is a clinical note feature. One hot encoding is used to represent the occurrence of each n-gram in the patient record. The n-grams are fit into the designed feature selection process and then integrated with the structured data for predictive analysis.

## 2.2 System framework

Figure 2 shows the system framework of our research, the structured data and clinical notes in our complete data set are first pre-processed by applying the steps mentioned in the previous section. Then, the complete data set is split into training and test data. The training set is used to create predictive models using logistic regression (LR), random forest (RF), support vector machines (SVM), Multi-Layer Perceptron (MLP), XGBoost, and Gradient Boosting (GB), respectively. We investigated three different approaches to integrate the structured data with clinical notes for sarcopenia detection. The first approach (referred to as “+Data-driven Clinical Notes Features (CNF)”) uses the data-driven feature selection method to first identify the relevant clinical note features with regards to the labels (control vs. sarcopenia) of the patients. In the data-driven feature selection approach, the note features are selected in a machine learning approach based on their impact on the fit of the model rather than on their conceptual connection to the outcome of interest. The selected clinical note features are treated as independent variables along with all structured data (without applying feature selection) to be fed into the learning models. The second approach (referred to as “+Data-driven All Features (AF) + Feature Selection (FS)”) uses the data-driven feature selection on both structured data and clinical note features to identify most important features. In the first and second approach, ANOVA F-test values were then calculated for each feature, and the elbow method<sup>27</sup> is used to determine the optimal number of features. Both “+Data-driven CNF” and “Data-driven AF + FS” do not involve intervention of human experts in feature selection (shown as orange arrow A in Figure 2). The third approach (referred to as “+Data-driven plus Knowledge-driven CNF”) integrates data-driven feature selection with the knowledge driven clinical note feature selection. After applying a feature selection to identify the relevant clinical note features, two or more domain experts go through the clinical note features to finalize the selection by limiting the features to those that are most relevant to sarcopenia from the expert knowledge point of view. The selected clinical note features are then fed into the learning models with the structured data (shown as yellow arrow B in Figure 2). The last approach (referred to as “+Augmented prediction using CNF”) does not feed the clinical note features into the learning models, but only adjust the prediction result based on the occurrences of the selected clinical note features identified by experts (show as blue arrow C in Figure 2). In this research, two clinicians (a full Professor and an Associate Professor) who are clinical and research faculty in an academic medical center in in the United States each with more than 10 years of experience in musculoskeletal disorders, identified the relevant clinical note features.

## 2.3 Data-driven and Knowledge-driven CNF

To reduce the number of features and improve the sensitivity of our model without sacrificing the specificity to improve the overall performance, we investigated this integration of data-driven and knowledge-driven approach. The ANOVA F-test is used as a feature selection method to calculate a F-test score for each clinical note feature. To determine the number of the relevant features to be selected, we applied elbow methods on the generated graph using F-test scores to select the top relevant features. Then, the clinical experts review the top scored features and select the most relevant clinical note features to be included in the model building.

## 2.4 Augmented prediction using CNF

The “Data-driven CNF”, “Data-driven AF + FS”, and the “Data-driven and Knowledge-driven CNF” approaches use both structured data and clinical note features to train the prediction models. The “+Augmented prediction using CNF” approach only uses structured data to train the models but then adjusts the final predictions based on presence or absence of the knowledge-driven features. The adjustment process works as: If the predicted probability of a patient being sarcopenia is less than the threshold that is determined by a prediction model, but more than the mean probability of all non-sarcopenic patients, the clinical note features are used to adjust the prediction result by increasing the probability to above the threshold when the patient possesses one or more of those clinical note features. On the other hand, if the predicted probability of sarcopenia is less than the mean of the non-sarcopenic patients, the presence of these text features is not used to recategorize the patient. For example, if the clinical notes contain one or more features, such as “cane”, etc., that indicate the patient could have sarcopenia, the prediction result is only adjusted upward if the model’s original prediction result using structured data indicates that patient’s probability of sarcopenia is higher than the mean probability of all non-sarcopenic patients. This prevents falsely adjusting upward the probability of those having otherwise low probability.

## 2.6 Evaluation Metrics and Statistics

AUC (Area Under the ROC Curve), sensitivity and specificity are used to evaluate the performance of the proposed model and compare baseline models that without using clinical note features. For each model, we selected the top parameter sets and thresholds that performed best on the validation set in terms of AUC, then calculated the AUC, sensitivity, and specificity on the test set. The equations of the sensitivity and specificity are given in Equation 1 and 2, where TP is true positive, FN is false negative, TN is true negative, and FP is false negative. To assess which model yields significant performance improvements over the models that are based on the structured data only, we conducted Mann-Whitney tests for the comparison. The statistical analyses were performed using Python version 3.9.0, and p-values < 0.05 were considered statistically significant.

$$\text{sensitivity} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{specificity} = \frac{TN}{TN+FP} \quad (2)$$

## 2.7 Experimental Setting

We split our dataset into training (70% of the complete data), validation (10% of the complete data), and a test set (20% of the complete data) with a stratified approach. Hence, training, validation, and test sets preserve the same proportions of sarcopenia patients as

observed in the original dataset. The validation set was used to fine-tune the parameters of the models, and the test set was used to evaluate the performance of all models. For all learning models, we use the top 10 best performed parameters to calculate the 95% confidence interval (CI) of AUC, sensitivity, and specificity values. To calculate confidence intervals, we first determine if the data follows a normal distribution by performing the Shapiro-Wilk and Kolmogorov-Smirnov tests. If both tests yield p-values above 0.05, we assume the data is normally distributed. Then, we calculate the 95% confidence interval using the mean and standard deviation. Otherwise, we create bootstrap samples from the original data, then calculate the 95% bootstrap confidence interval.

### 3. Results

#### 3.1 Statistics of the dataset

After implementing the preprocessing steps for structured data and clinical notes, the ultimate dataset comprised 1304 patients, each associated with a total of 344 structured variables. These variables encompassed 205 distinct diagnosis codes, 67 laboratory tests, 71 medications, and Body Mass Index (BMI). Among the participants included, 1055 did not meet any criteria for sarcopenia (controls), while 249 met at least one of the sarcopenia test thresholds as in the prior publication<sup>21</sup>. Demographic information and number of structured data and clinical notes of the dataset are summarized in Table 1. The sarcopenia group was older, more likely to be male or African American, and had higher BMI than the control group.

#### 3.2 Selected Clinical Note Features

By applying the data-driven clinical note feature selection process, we first generated 252,546 word-based n-grams including 1-, 2-, and 3-grams. The ANOVA F-test values were then calculated for each gram. We applied the elbow method<sup>27</sup> to select the top 189 n-grams (listed in supplemental document) that have the highest F-test values. Figure 1 shows the threshold determined by the elbow method. The selected n-grams included “spine”, “bilateral leg”, “marginal osteophytes”, and “positive musculoskeletal pain”, etc. However, most n-grams were considered likely non-specific to sarcopenia, though they could impact performance on sarcopenia tests. From the top ranked 189 n-grams, experts identified five n-grams that may be more likely to indicate whether the patient tests as sarcopenia based on their knowledge and literature<sup>28–31</sup>: “steady”, “unsteady”, “cane”, “muscle tenderness”, and “walker”. These n-grams either could be used to describe the walking behavior or needs of a patient or describe a muscle symptom that are relevant to sarcopenia. “Sarcopenia” as a term was not found in these patients’ clinical notes; hence it is not included.

#### 3.3 Performance Evaluation and Comparison

Table 2, 3, and 4 show the AUC values, sensitivity, and specificity of baseline models (using structured data only) and four different approaches to integrating clinical note features with the structured data. The structured data only model includes all diagnosis, medication, laboratory test, and BMI variables without applying any feature selection. The final number of features used to build the learning models are: 344, 533, 45, 349 and 344, respectively for the “Structured data only”, “+ Data-driven CNF”, “+ Data-driven AF + FS”, “+ Data-

driven and Knowledge-driven CNF” and “+ Augmented prediction using CNF” approaches. The results indicate that for all learning models, most methods of incorporating clinical note features led to improved outcomes. Four models - Logistic regression (LG), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Gradient Boosting (GB) work better when “+ Augmented prediction using CNF” approach is applied. Whereas XGBoost achieves a better result when adding the clinical features using the data-driven approach. When adding the clinical note features produced by combined data-driven and knowledge-driven approaches, Random Forest gained a better result than its baseline. The statistical t-test shows that the results gained by LR and SVM are significantly better than the rest of the approaches when “+ Augmented prediction using CNF” is applied. It is worth noting that although “+ Data-driven AF + FS” has only 45 input features, the performance is still better than the “structured data only” approach with most of the learning models. This demonstrates the importance of adding clinical note features for sarcopenia prediction.

The results indicate that, for every machine learning algorithm, there is at least one approach incorporating clinical note features that enhances sensitivity compared to the baseline, though for some models the sensitivity decreased, especially using the data driven CNF. For LR, SVM, MLP, and GB, the sensitivity increased 3.7% when adjusted results using CNF approach is used. For RF and XGBoost the sensitivity increased more than 8% when “+Data-driven CNF” approach is used. It is noticed that the “+Data-driven CNF” approach can significantly increase the specificities gained by LR, SVM and MLP, but sacrifice the sensitivity values. Whereas augmenting the prediction results using CNF approach can increase the sensitivities without sacrificing specificities. Hence, the “+ Augmented prediction using CNF” approach performs better overall.

Figure 4 shows the calibration slopes for the training and test datasets using the “+ Augmented prediction using CNF” approach. The best Brier scores gained on training data were 0.127 when MLP and LR were used, respectively. The best Brier score gained on test data were 0.137 and 0.139 when XGBoost and LR were used, respectively.

### 3.4 Case study

To further understand the impact of adding clinical note features in different models, we analyzed two different scenarios in this case study section. The case study is done using the local explanation capability of the Shapley Additive exPlanations (SHAP) method<sup>32</sup>. SHAP has been used in applying ML models for various clinical outcomes or disease prediction<sup>33–35</sup>. The objective of SHAP is to explain the prediction  $f(x)$  for a given instance  $x$  by determining the proportionate contribution of each feature value to the particular outcome. SHAP has different explainers for different machine learning models. For example, Linear Explainer in the SHAP library can be used for linear models, whereas the Tree Explainer in the SHAP library can be used for tree-based models (XGBoost, Random Forest, etc.). To visualize the local explanation, we used waterfall plot. The base of the waterfall plot has an expected value of the model output. The waterfall plot organizes features in descending order of importance, with the length of the bars reflecting the proportional feature contribution, and red or blue of the bars reflecting the positive or negative impact of individual features. The top 10 important features are shown in the waterfall plots.

The first case (shown in Figure 5) demonstrates one patient who was a 56-year-old female and sarcopenic based on the test measurement criteria (ground truth) for sarcopenia. The waterfall plot (Figure 5a) shows that the diagnoses codes in category “symptoms and signs involving the nervous and musculoskeletal systems”, “slipping, tripping, stumbling and falls”, “injuries to the head” are important positive factors predicting sarcopenia. The literature<sup>36</sup> shows that the sympathetic nervous system affects skeletal muscle composition and influences the development of sarcopenia. Patients with sarcopenia also have a higher risk of falling, tripping and slipping<sup>37,38</sup>, which can result in injuries to head<sup>39</sup> etc. Sarcopenia patients often have abnormal findings in functional studies<sup>40,41</sup> and take medications for musculoskeletal symptoms<sup>42</sup>. When only structured data is used, the expected model output is lower than the actual output. Hence, the overall prediction of this case by logistic regression model is non-sarcopenic. After applying the “+Augmented prediction using CNF” approach, the prediction is changed to sarcopenic, because of the presence of two relevant clinical note features “cane” and “unsteady” in the clinical notes. The snippet of the clinical note is shown in Figure 5b.

The second case (shown in Figure 6) demonstrates a patient who was who was a 75-year-old male and sarcopenic based on the test measurement criteria (ground truth) for sarcopenia. The diagnoses “diabetes mellitus”, “Overweight, obesity and other hyperalimentation”, “other forms of heart diseases”, “metabolic disorders” and “general symptoms and signs” are positive factors predicting sarcopenia. The literature shows<sup>43</sup> that diabetes and sarcopenia have bidirectional relations. The reduction of muscle mass is strongly associated with insulin resistance and the development of metabolic syndrome<sup>43,44</sup>. Research<sup>45,46</sup> also shows a high prevalence of sarcopenia among older adults with coronary heart disease or complex congenital heart disease. However, without adding the clinical note features, the random forest (RF) model predicted it as non-sarcopenic. After applying the “+data-driven and knowledge-driven CNF” approach, the case is correctly predicted as sarcopenic. The actual model output is larger than the expected output. It shows that adding the clinical note feature “walker” has a positive impact on the model. The snippet of the clinical note is shown in Figure 6b.

The third case (shown in Figure 7) demonstrates a patient who was a 25-year-old female and sarcopenic based on the test measurement criteria (ground truth) for sarcopenia. The diagnoses “aplastic anemias and other bone marrow failure syndromes”, and “inflammatory diseases of female pelvic organs” were positive factors predicting sarcopenia. The literature shows<sup>47,48</sup> that sarcopenia can develop in patients with aplastic anemias and among patients who require haematopoietic cell transplantation (HSCT). However, without adding the clinical note features, the support vector machine (SVM) model predicted it as non-sarcopenic. After applying the “+Augmented prediction using CNF” approach, the case is predicted as sarcopenic. It shows that adding clinical note feature “muscle tenderness” has a positive impact on the model. The snippet of the clinical note is shown in Figure 7b.

#### 4. Discussion

Timely identification of sarcopenia is essential, as it enables the application of interventions to impede the advancement of the condition, enhance the quality of life for those

affected, and decrease healthcare expenses. However, sarcopenia often goes undiagnosed and undocumented. Recent research shows that it is feasible to utilize specifically collected medical and muscle evaluation data<sup>49–51</sup>, or physical activity data to detect sarcopenia<sup>52</sup>. Very few research studies have investigated using EHR data for sarcopenia detection, so physicians can efficiently identify patients at risk. Most existing approaches<sup>52,53</sup> for predicting sarcopenia rely on data from the National Health and Nutrition Examination Survey (NHANES), which includes a broader range of patients and prospectively collected specific additional data not typically available in the EHR. Some of these approaches show better performances in terms of AUC (0.831) than ours (0.739), but likely reflect the different data sources. Our research lays a foundation for clinical decision support components which could be implemented within the EHR by utilizing the EHR data itself as a predictor. We previously demonstrated that using structured EHR data machine learning models can be built to identify predictors for sarcopenia. However, information written in the clinical notes can provide important indicators of early sarcopenia that might not rise to the level of detection through diagnosis codes or other diagnostic evaluations. In this study, we built on our prior machine learning work using structured data and we evaluated the impact of different approaches to integrate the clinical note features with the structured data from the EHR to improve the detection of sarcopenia prediction using various machine learning approaches. The performances show that when machine learning models are used with the “+Augmented prediction using CNF” approach, some model can significantly improve the sensitivity without reducing the specificity of sarcopenia detection over the baseline models that had only used structured data. When “+Data-driven CNF” approach is used, the specificity improves with a drop in sensitivity. The case studies also show that adding clinical notes can also enhance the interpretation of the models by linking the information written in the clinical notes with the structured data. We envision in the future creating a clinical decision support component using the “+Augmented Prediction Using CNF” approach on a large patient cohort, making it generalizable within the EHR system. This method allows human experts to validate system-ranked clinical note features before integrating them into predictive analysis. Additionally, incorporating SHAP visualization can enhance the model’s clinical utility by providing a case-based visualization tool.

Extracting disease relevant features from a large number of clinical notes can be challenging, especially when many different note features may be important to a disease. In the current research we developed a natural language processing (NLP) pipeline to first identify the sentences or chunks of clinical notes that might contain information that falls into a set of semantic types, such as ‘Activity’, and ‘Sign or Symptom’ etc., using UMLS MetaMap<sup>26</sup>, then we extracted n-grams and ranked them using feature selection strategy. Although the data-driven CNF approach improved the performances over the baseline approach, the extracted features might not all be supported by literature or might need further investigation. For example, the features “degenerative disc disease”, “lost balance”, “edema”, etc. have shown positive impact on sarcopenia detection in the data driven models, resulting in their inclusion as features in the data-driven CNF approaches. However, these features, and others that may predict sarcopenia, are not conceptually specific to sarcopenia, and may occur in clinical notes due to transient conditions or as the result of other conditions. These features might be study cohort dependent and might not be generalizable

to a larger population. However, the knowledge driven features “cane”, “walker” etc., are interpretable by experts as being likely to explain performance on sarcopenia tests, and thus be relevant to patients who might have sarcopenia. Hence, it is necessary to combine data-driven and knowledge-driven approaches in clinical research. Human experts can contribute to selection at the feature validation stage to enhance the model’s generalizability and mitigate biases introduced by the data sources and enhance the interpretability of the ML models. It is worth noting that to avoid individual human biases, two or more human experts are needed to select and finalize the features to be included in the knowledge-driven approach.

It is worth noting that not all the clinical notes of a patient contain information that is relevant to the disease. For example, diagnostic imaging reports often do not have any description of the patients’ walking or activity behavior description that might be relevant to sarcopenia. While radiologic imaging may be useful to quantitate muscle size or other muscle characteristics, clinical imaging reports (usually done for other purposes) did not usually contain such information to characterize sarcopenia. Hence, including clinical reports that contain relevant information is crucial. On the other hand, physician notes may often not record all possible sarcopenia relevant information in order for the system to integrate this information for analysis.

#### 4.1 Limitations of the study

Our research objective was to demonstrate the performances of the ML models on sarcopenia detection using EHR with and without text from clinical notes. We recognize the limitations of this study which include the following. The study cohort is not very large, although it includes various age groups and patients with different co-morbidities. Hence, further validation needs to be done to investigate which approach can be generalized to a larger population. The traditional NLP method is designed to process the clinical notes to identify candidate features. A more sophisticated NLP algorithm based on the large language models could be used to identify the candidate clinical note features in the future to possibly improve the performance of the data-driven approaches. However, there are patient privacy and legal concerns that would be raised if assessing clinical notes using external large language model platforms, as these notes can be difficult to completely de-identify. Thus, we could not apply those models to our data. Furthermore, many terms that may indicate sarcopenia (such as “muscle weakness”) are very nonspecific and many patients without sarcopenia report such a symptom to be documented in clinical notes, due to the subjectivity of the symptom, or the symptom being a focal or temporary condition associated with an illness or injury that may not reflect generalized sarcopenia or later performance on formal testing. While these participants had engaged in formal research measurements enabling the categorization as sarcopenic, they did not undergo a formal focused clinical evaluation to determine who should and should not have sarcopenia based on clinical history, symptoms, or other assessments in their EHR. A formal epidemiology study of sarcopenia could conduct such detailed questioning enabling identification of other predictors of sarcopenia. However, that would not be able to reflect the current clinical problem of underdiagnosis of sarcopenia and a lack of specific features found in EHR, which our study was conducted to address. Finally, the involvement of human experts in

validating the selection of clinical note features might lead to disagreements on feature determination. To resolve such conflicts, multiple human experts are necessary, which could add to their workload. However, we believe that human-in-the-loop AI will be essential in the future to validate AI outputs and continually improve AI systems.

## 5. Conclusion

The current research investigated four different approaches to integrate features extracted from the narrative clinical notes with the structured data in the EHR for sarcopenia detection. The results show that using data-driven and knowledge-driven approaches together can improve performance and avoid biases that can be introduced by the data-driven approach. Future studies are needed to assess larger cohorts incorporating validated sarcopenia patients with detailed EHR data availability to enable a generalizable model to be evaluated in real-world clinical settings. The proposal framework can be applied to diseases like sarcopenia that need early detection by including key factors written in the narrative clinical notes.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Funding support:

This work was supported in part by funding from the NIH by NIAMS (P30AR072581 and R01AR077273) and by NIDDK (K23DK102824) and by NCATS (UL1TR002529). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The sponsor was not involved in the study conduct or writing or approving the manuscript.

## References

1. Studenski SA, Peters KW, Alley DE, et al. The FNIH sarcopenia project: rationale, study description, conference recommendations, and final estimates. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*. 2014;69(5):547–558.
2. Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age and ageing*. 2019;48(1):16–31. [PubMed: 30312372]
3. Artaud F, Singh-Manoux A, Dugravot A, Tzourio C, Elbaz A. Decline in fast gait speed as a predictor of disability in older adults. *Journal of the American Geriatrics Society*. 2015;63(6):1129–1136. [PubMed: 26096387]
4. Heiland EG, Welmer AK, Wang R, et al. Association of mobility limitations with incident disability among older adults: a population-based study. *Age and ageing*. 2016;45(6):812–819. [PubMed: 27126329]
5. Beavers KM, Beavers DP, Houston DK, et al. Associations between body composition and gait-speed decline: results from the Health, Aging, and Body Composition study. *The American journal of clinical nutrition*. 2013;97(3):552–560. [PubMed: 23364001]
6. Reinders I, Murphy RA, Koster A, et al. Muscle quality and muscle fat infiltration in relation to incident mobility disability and gait speed decline: the Age, Gene/Environment Susceptibility-Reykjavik Study. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*. 2015;70(8):1030–1036.
7. Beaudart C, Zaaria M, Pasleau F, Reginster JY, Bruyère O. Health outcomes of sarcopenia: a systematic review and meta-analysis. *PLoS one*. 2017;12(1):e0169548. [PubMed: 28095426]

8. Hiraoka A, Tamura R, Oka M, et al. Prediction of risk of falls based on handgrip strength in chronic liver disease patients living independently. *Hepatology research*. 2019;49(7):823–829. [PubMed: 30770617]
9. Landi F, Liperoti R, Russo A, et al. Sarcopenia as a risk factor for falls in elderly individuals: results from the iLSIRENTE study. *Clinical nutrition*. 2012;31(5):652–658. [PubMed: 22414775]
10. Sim M, Prince R, Scott D, et al. Utility of four sarcopenia criteria for the prediction of falls-related hospitalization in older Australian women. *Osteoporosis international*. 2019;30:167–176. [PubMed: 30456572]
11. Tanaka T, Son BK, Lyu W, Iijima K. Impact of social engagement on the development of sarcopenia among community-dwelling older adults: A Kashiwa cohort study. *Geriatrics & Gerontology International*. 2022;22(5):384–391. [PubMed: 35322539]
12. Tagliafico AS, Bignotti B, Torri L, Rossi F. Sarcopenia: how to measure, when and why. *La radiologia medica*. 2022;127(3):228–237. [PubMed: 35041137]
13. Arango-Lopera V, Arroyo P, Gutiérrez-Robledo LM, Pérez-Zepeda MU, Cesari M. Mortality as an adverse outcome of sarcopenia. *The journal of nutrition, health & aging*. 2013;17(3):259–262.
14. Landi F, Cruz-Jentoft AJ, Liperoti R, et al. Sarcopenia and mortality risk in frail older persons aged 80 years and older: results from iLSIRENTE study. *Age and ageing*. 2013;42(2):203–209. [PubMed: 23321202]
15. De Buyser SL, Petrovic M, Taes YE, et al. Validation of the FNIH sarcopenia criteria and SOF frailty index as predictors of long-term mortality in ambulatory older men. *Age and ageing*. 2016;45(5):602–608. [PubMed: 27126327]
16. Steffl M, Sima J, Shiells K, Holmerova I. The increase in health care costs associated with muscle weakness in older people without long-term illnesses in the Czech Republic: results from the Survey of Health, Ageing and Retirement in Europe (SHARE). *Clinical interventions in aging*. Published online 2017:2003–2007. [PubMed: 29225462]
17. Goates S, Du K, Arensberg M, Gaillard T, Guralnik J, Pereira SL. Economic impact of hospitalizations in US adults with sarcopenia. *The Journal of frailty & aging*. 2019;8:93–99. [PubMed: 30997923]
18. Reijnierse EM, De Van Der Schueren MA, Trappenburg MC, Doves M, Meskers CG, Maier AB. Lack of knowledge and availability of diagnostic equipment could hinder the diagnosis of sarcopenia and its management. *PloS one*. 2017;12(10):e0185837. [PubMed: 28968456]
19. Vellas B, Fielding R, Bens C, et al. Implications of ICD-10 for sarcopenia clinical practice and clinical trials: report by the international conference on frailty and sarcopenia research task force. *The Journal of frailty & aging*. 2018;7:2–9. [PubMed: 29412436]
20. Moorthi RN, Liu Z, El-Azab SA, et al. Sarcopenia, frailty and cachexia patients detected in a multisystem electronic health record database. *BMC Musculoskeletal Disorders*. 2020;21:1–8.
21. Luo X, Ding H, Broyles A, Warden SJ, Moorthi RN, Imel EA. Using machine learning to detect sarcopenia from electronic health records. *Digital Health*. 2023;9:20552076231197098.
22. McLean RR, Shardell MD, Alley DE, et al. Criteria for clinically relevant weakness and low lean mass and their longitudinal association with incident mobility impairment and mortality: the foundation for the National Institutes of Health (FNIH) sarcopenia project. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*. 2014;69(5):576–583.
23. Warden SJ, Kemp AC, Liu Z, Moe SM. Tester and testing procedure influence clinically determined gait speed. *Gait & posture*. 2019;74:83–86. [PubMed: 31491564]
24. Warden SJ, Liu Z, Moe SM. Sex-and age-specific centile curves and downloadable calculator for clinical muscle strength tests to identify probable sarcopenia. *Physical Therapy*. 2022;102(3):pzab299.
25. Stausberg J, Lehmann N, Kaczmarek D, Stein M. Reliability of diagnoses coding with ICD-10. *International journal of medical informatics*. 2008;77(1):50–57. [PubMed: 17185030]
26. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. Published online 2001:17–21. [PubMed: 11825149]
27. Antunes M, Ribeiro J, Gomes D, Aguiar RL. Knee/elbow point estimation through thresholding. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE; 2018:413–419.

28. Menant J, Weber F, Lo J, et al. Strength measures are better than muscle mass measures in predicting health-related outcomes in older people: time to abandon the term sarcopenia? *Osteoporosis international*. 2017;28:59–70. [PubMed: 27394415]
29. Abellan van Kan G. Epidemiology and consequences of sarcopenia. *JNHA-The Journal of Nutrition, Health and Aging*. 2009;13:708–712.
30. Shoemaker M, Mustad V, Pereira S, et al. Metabolic Differences During Submaximal, Steady-State Aerobic Exercise between Sarcopenic and Non-Sarcopenic Older Adults. *Current Developments in Nutrition*. 2021;5:524.
31. Bravo-José P, Moreno E, Espert M, Romeu M, Martínez P, Navarro C. Prevalence of sarcopenia and associated factors in institutionalised older adult patients. *Clinical nutrition ESPEN*. 2018;27:113–119. [PubMed: 30144883]
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.
33. Wang K, Tian J, Zheng C, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Computers in Biology and Medicine*. 2021;137:104813. [PubMed: 34481185]
34. Raihan MJ, Khan MAM, Kee SH, Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*. 2023;13(1):6263. [PubMed: 37069256]
35. Lu S, Chen R, Wei W, Belovsky M, Lu X. Understanding heart failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions. In: *AMIA Annual Symposium Proceedings*. Vol 2021. American Medical Informatics Association; 2021:813. [PubMed: 35308970]
36. Delbono O, Rodrigues ACZ, Bonilla HJ, Messi ML. The emerging role of the sympathetic nervous system in skeletal muscle motor innervation and sarcopenia. *Ageing research reviews*. 2021;67:101305. [PubMed: 33610815]
37. Fadem SZ. *Understanding and Preventing Falls: A Guide to Reducing Your Risks*. Springer Nature; 2023.
38. Martin FC. Frailty, sarcopenia, falls and fractures. *Orthogeriatrics*. Published online 2017:47–61.
39. Correa RGP, Pivovarsky MLF, Santos G da S, Gomes ARS, Borba VZC. Factors that cause women with osteoporosis to fall. *Archives of endocrinology and metabolism*. 2023;67(4):e000578. [PubMed: 37252691]
40. Ueshima J, Maeda K, Shimizu A, et al. Diagnostic accuracy of sarcopenia by “possible sarcopenia” premiered by the Asian Working Group for Sarcopenia 2019 definition. *Archives of gerontology and geriatrics*. 2021;97:104484. [PubMed: 34298259]
41. Boutin RD, Yao L, Canter RJ, Lenchik L. Sarcopenia: current concepts and imaging implications. *American Journal of Roentgenology*. 2015;205(3):W255–W266. [PubMed: 26102307]
42. Iolascon G, Moretti A, De Sire A, Liguori S, Toro G, Gimigliano F. Pharmacological therapy of sarcopenia: past, present and future. *Clinical Cases in Mineral & Bone Metabolism*. 2018;15(3).
43. Mesinovic J, Zengin A, De Courten B, Ebeling PR, Scott D. Sarcopenia and type 2 diabetes mellitus: a bidirectional relationship. *Diabetes, metabolic syndrome and obesity: targets and therapy*. Published online 2019:1057–1072. [PubMed: 31372016]
44. Nishikawa H, Asai A, Fukunishi S, Nishiguchi S, Higuchi K. Metabolic syndrome and sarcopenia. *Nutrients*. 2021;13(10):3519. [PubMed: 34684520]
45. Zhang N, Zhu WL, Liu XH, et al. Prevalence and prognostic implications of sarcopenia in older patients with coronary heart disease. *Journal of geriatric cardiology: JGC*. 2019;16(10):756. [PubMed: 31700515]
46. Sandberg C, Johansson K, Christersson C, Hlebowicz J, Thilén U, Johansson B. Sarcopenia is common in adults with complex congenital heart disease. *International journal of cardiology*. 2019;296:57–62. [PubMed: 31230936]
47. Chen D, Yuan Z, Guo Y, et al. The evolution and impact of sarcopenia in severe aplastic anaemia survivors following allogeneic haematopoietic cell transplantation. *Journal of Cachexia, Sarcopenia and Muscle*. Published online 2024.

48. Chen D, Yuan Z, Guo Y, et al. Prognostic impact of quantifying sarcopenia and adipopenia by chest CT in severe aplastic anemia patients treated with allogeneic hematopoietic stem cell transplantation. *Academic Radiology*. 2023;30(9):1936–1945. [PubMed: 36379814]
49. Agnes T, Vishal K, others. Regression model for the prediction of risk of sarcopenia among older adults. *Muscles, Ligaments & Tendons Journal (MLTJ)*. 2019;9(3).
50. Liao H, Yang Y, Zeng Y, et al. Use machine learning to help identify possible sarcopenia cases in maintenance hemodialysis patients. *BMC nephrology*. 2023;24(1):1–17. [PubMed: 36597041]
51. Turimov Mustapoevich D, Kim W. Machine Learning Applications in Sarcopenia Detection and Management: A Comprehensive Survey. In: *Healthcare*. Vol 11. MDPI; 2023:2483. [PubMed: 37761680]
52. Seok M, Kim W. Sarcopenia Prediction for Elderly People Using Machine Learning: A Case Study on Physical Activity. In: *Healthcare*. Vol 11. MDPI; 2023:1334. [PubMed: 37174876]
53. Seok M, Kim W, Kim J. Machine learning for sarcopenia prediction in the elderly using socioeconomic, infrastructure, and quality-of-life data. In: *Healthcare*. Vol 11. MDPI; 2023:2881. [PubMed: 37958025]

### Summary Table

- It is crucial to combine both the structured data with the unstructured data from the Electronic Health Records for clinical decision support and disease detection.
- We investigated various approaches to integrate the clinical text notes into models for sarcopenia detection.
- We evaluated and developed a model to integrate the data-driven and knowledge-driven approaches to incorporate text data.
- Further research needs to be done on larger cohorts and to evaluate the feasibility of developing a clinical decision support component for the EHR

..... She uses a **walker** most of the time. ...

*Medical device*

The patient is clearly **disabled in her activities of daily living** because of

*Daily or Recreational Activity*

this **right hip pain** and is incapable of caring lot of activities that she used to be able to....

*Sign or Symptom .*

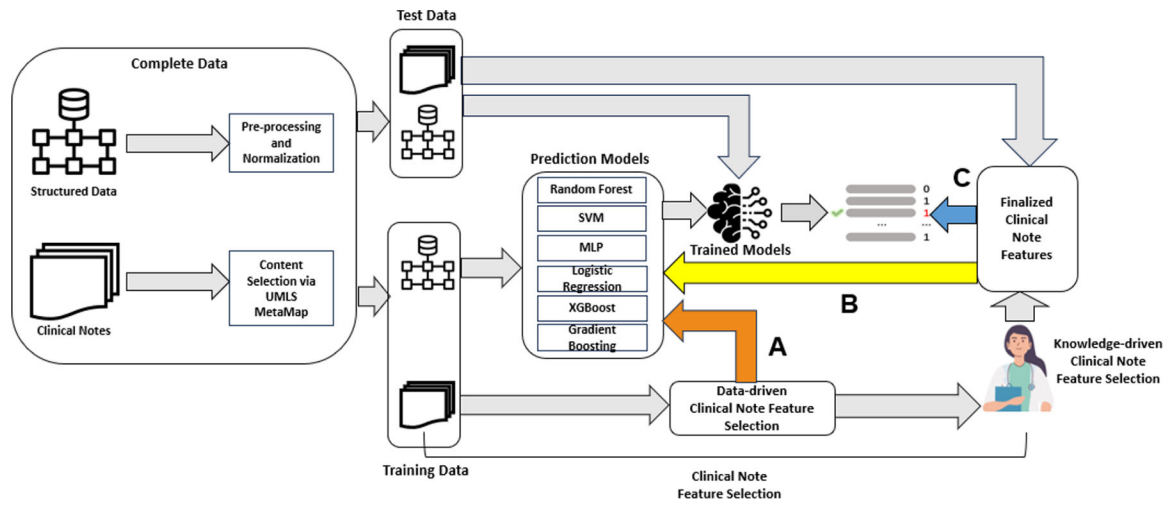
..... Currently, the pain is getting worse when she **walks** long distances and when she

*Daily or Recreational Activity*

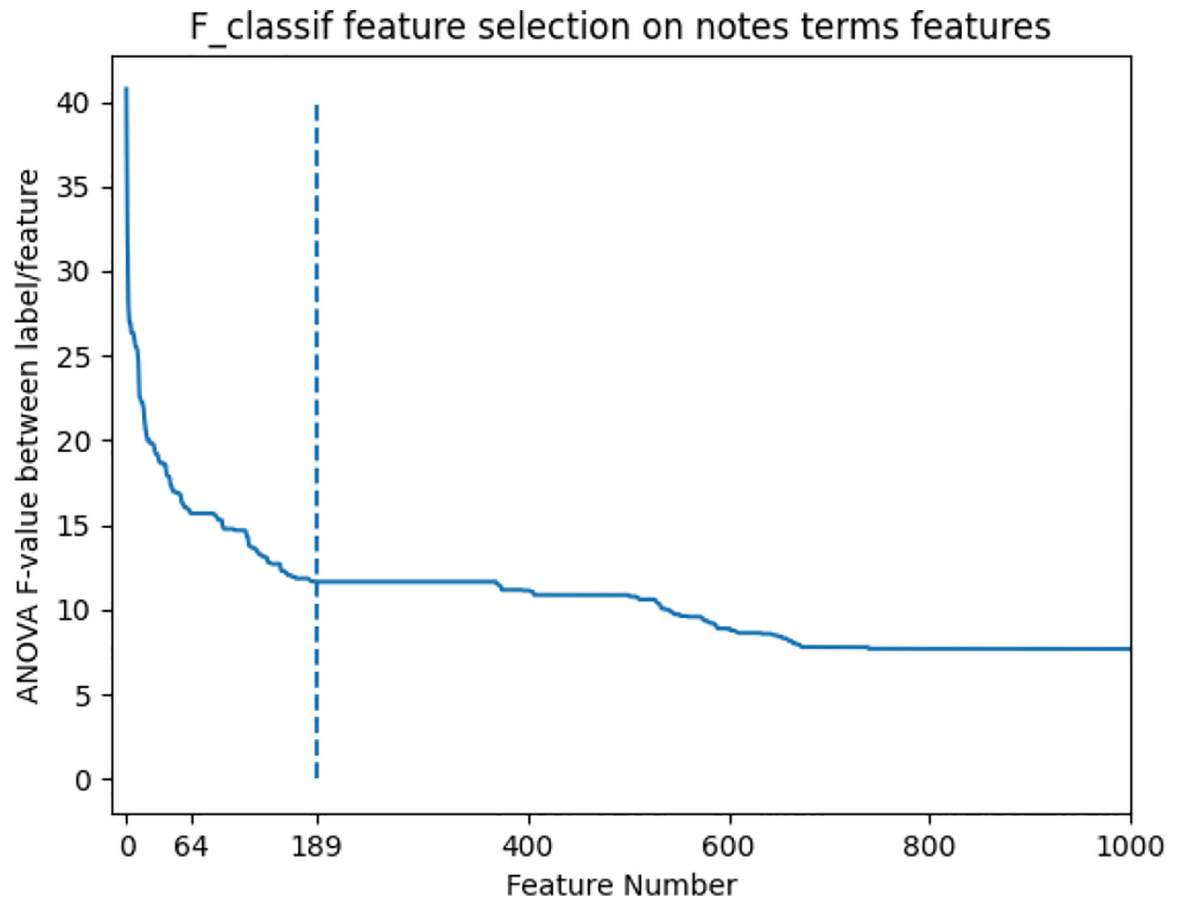
goes up and down stairs. ....

**Figure 1.**

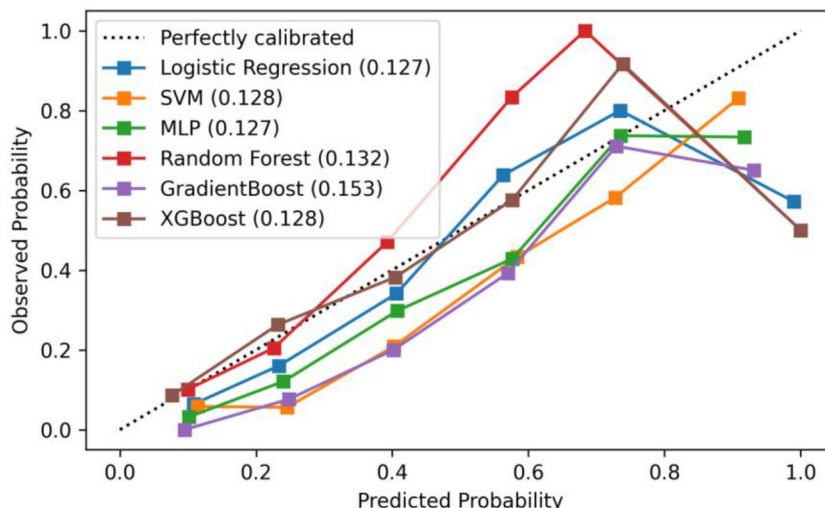
Sentences that are selected using UMLS MetaMap



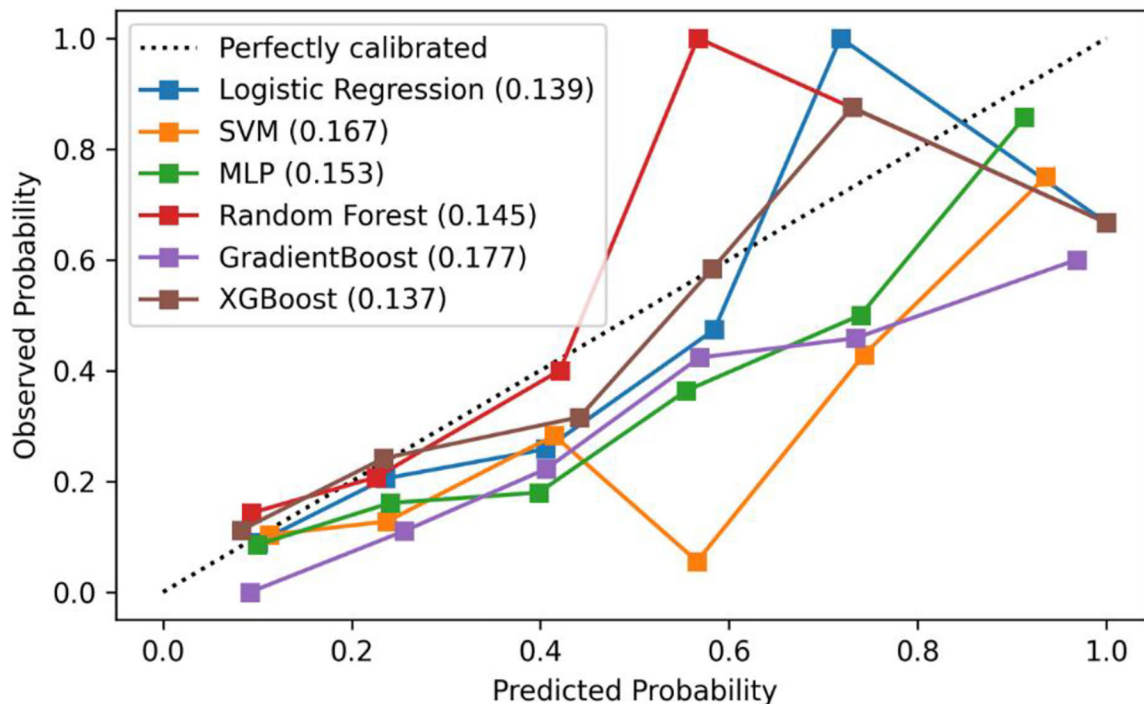
**Figure 2.**  
System Framework



**Figure 3.**  
Elbow Method to Select Top Clinical Note Features



a Brier Score Calibration Curves for training dataset



b Brier Score Calibration Curves for test dataset

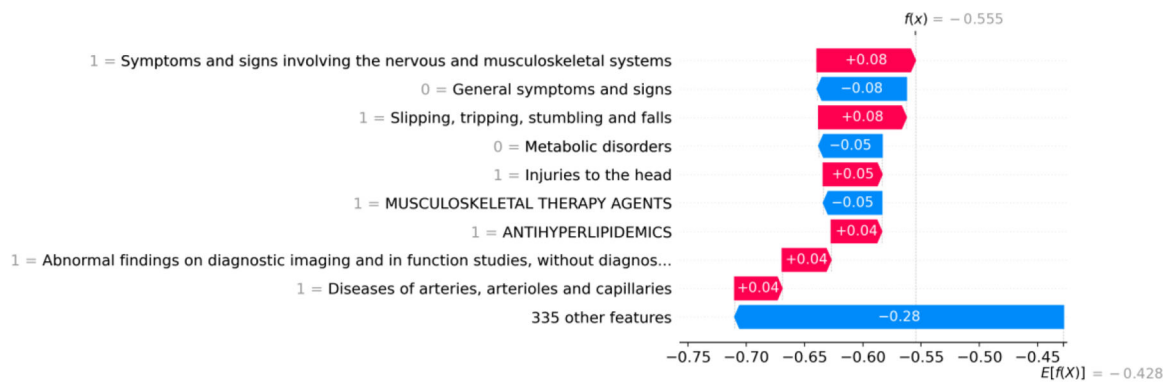
Figure 4. Brier Score Calibration Curves using the “+ Augmented prediction using CNF” approach.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



a SHAP Waterfall plots for Case 1

.....

Since his dx of MS he has had trouble/weakness in the RLE so ambulates with a cane.

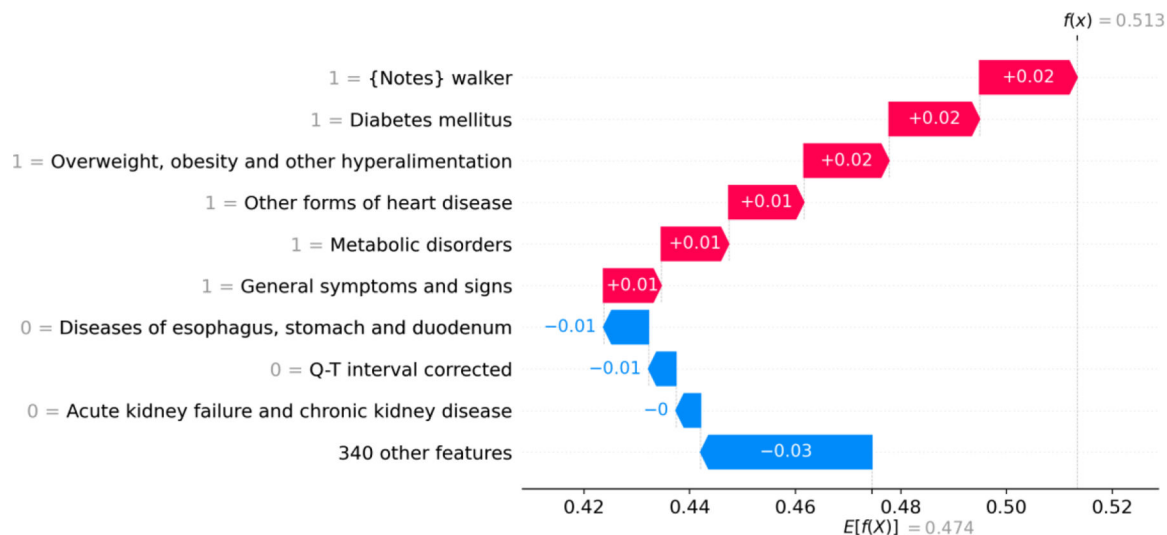
.....

Station and Gait: ambulates w. cane in right hand, mildly unsteady with casual gait right foot drop and circumduction.

.....

b Clinical note snippet of Case 1

Figure 5.  
Case Study 1

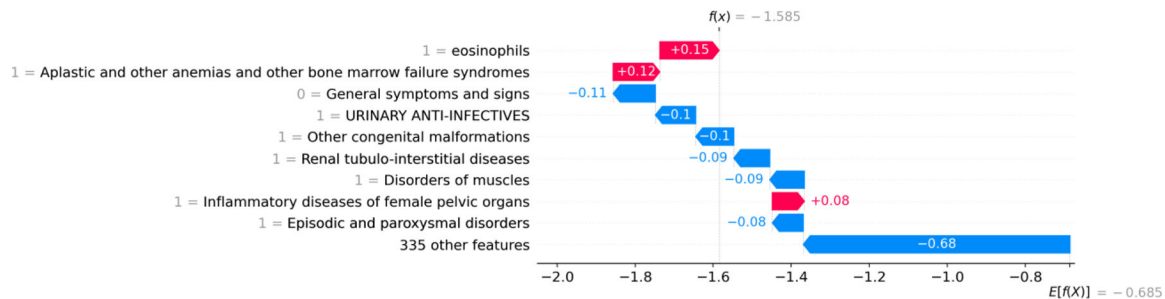


a SHAP Waterfall plots for Case 2

..... She does endorse ongoing fatigue and ambulates with a **walker**, but has been going to the YMCA for exercise three times a week. She continues to have blurry vision on the left eye that she attributes to cataract surgery. ....

b Clinical note snippet of Case 2

**Figure 6.**  
Case Study 2



a SHAP Waterfall plots for Case 3

.... Reports that she underwent corrective surgery in 2015. Reports over past 2 months has felt some pressure in her chest. Reports feels muscle tenderness since surgery .....

b Clinical note snippet of Case 3

Figure 7.  
Case Study 3

**Table 1.**Study cohort characteristics<sup>1</sup>

	<b>Total</b>	<b>Sarcopenia</b>	<b>Control</b>	<b>p value</b>
Number of Patients	1304	249	1055	
<b>Demographics</b>				
Age	53.32 (15.2)	58.36 (16.1)	52.13 (14.7)	<0.01
Male	263 (20.2%)	72 (28.9%)	191 (18.1%)	<0.01
Female	1041 (79.8%)	177 (71.1%)	864 (81.9%)	<0.01
White	1098 (84.2%)	192 (77.1%)	906 (85.9%)	<0.01
African American	111 (8.5%)	36 (14.5%)	75 (7.1%)	<0.01
Asian	85 (6.5%)	19 (7.6%)	66 (6.3%)	0.5172
Other	10 (0.8%)	2 (0.8%)	8 (0.8%)	0.7408
BMI, kg/m <sup>2</sup>	28.47 (6.9)	30.74 (8.54)	27.94 (6.33)	<0.01
<b>Chronic Diseases</b>				
Diabetes	180 (13.8%)	71 (28.51%)	109 (10.33%)	<0.01
Dementia	7 (0.54%)	5 (2.01%)	2 (0.19%)	<0.01
Cerebrovascular disease	52 (3.99%)	19 (7.63%)	33 (3.13%)	<0.01
Hemiplegia or paraplegia	6 (0.46%)	4 (1.61%)	2 (0.19%)	<0.05
Congestive heart failure	56 (4.29%)	30 (12.05%)	26 (2.46%)	<0.01
Peripheral vascular disease	41 (3.14%)	20 (8.03%)	21 (1.99%)	<0.01
Renal disease	78 (5.98%)	39 (15.66%)	39 (3.70%)	<0.01
Any malignancy	150 (11.50%)	45 (18.07%)	105 (9.95%)	<0.01
Osteoporosis	135 (10.35%)	30 (12.05%)	105 (9.95%)	0.39
# structured variables per patient	73.85 (31.46)	92.94 (40.61)	69.34 (27)	<0.01
# clinical notes per patient	6.41 (11.32) (out of 1304)	11.97 (17.82) (out of 249)	5.10 (8.63) (out of 1055)	<0.01

<sup>1</sup>Continuous variables are shown as mean (SD), and frequencies as n (%). The Mann-Whitney test was applied to the continuous variables, and chi<sup>2</sup> test was applied to the discrete variables.

**Table 2**Performance Comparison of AUC<sup>2</sup>

Models	Structured data only		+ Data-driven CNF		+ Data-driven AF + FS		+ Data-driven and Knowledge-driven CNF		+ Augmented prediction using CNF	
	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)
LR	71.59	(71.51, 71.66)	73.06**	(73.00, 73.11)	72.13**	(71.99, 72.27)	71.67**	(71.61, 71.73)	<b>73.93**</b>	<b>(73.64, 74.22)</b>
SVM	71.17	(69.28, 73.07)	72.14	(69.70, 74.57)	72.29*	(70.91, 73.66)	72.18*	(71.17, 73.18)	<b>73.87**</b>	<b>(72.18, 75.56)</b>
MLP	71.48	(71.00, 71.97)	72.34**	(71.64, 73.04)	72.49**	(72.22, 72.76)	71.96**	(71.55, 72.36)	<b>73.42**</b>	<b>(72.77, 74.06)</b>
RF	69.09	(67.97, 70.22)	71.21**	(70.30, 72.13)	71.59**	(70.23, 72.95)	<b>71.96**</b>	<b>(71.85, 72.06)</b>	71.30**	(70.24, 72.37)
GB	69.64	(69.30, 69.98)	62.55**	(61.81, 63.29)	64.41**	(62.17, 66.66)	70.45**	(70.20, 70.70)	<b>71.33**</b>	<b>(71.07, 71.58)</b>
XGBoost	70.19	(69.18, 71.19)	<b>73.42**</b>	<b>(73.12, 73.71)</b>	71.95**	(71.88, 72.02)	71.99**	(71.80, 72.18)	72.09**	(71.35, 72.83)

The \* represents p-value < 0.05 and \*\* represents p-value < 0.01 to show the significance of the improved performances over the baseline models. The best results of each learning model are highlighted as bold.

**Table 3**Performance Comparison of Sensitivity<sup>3</sup>

Models	Structured data only		+ Data-driven CNF		+ Data-driven AF + FS		+ Data-driven and Knowledge-driven CNF		+ Augmented prediction using CNF	
	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)
LR	65.74	(54.01, 77.47)	64.81	(64.81, 64.81)	61.11*	(61.11, 61.11)	69.26	(54.02, 84.50)	<b>69.44</b>	<b>(57.71, 81.18)</b>
SVM	67.59	(58.67, 76.52)	57.04	(43.67, 70.40)	60.55*	(48.62, 72.49)	69.44	(58.05, 80.84)	<b>71.30</b>	<b>(62.37, 80.22)</b>
MLP	67.04	(62.8, 71.27)	62.22	(54.40, 70.04)	54.63**	(52.19, 57.07)	67.22	(64.39, 70.06)	<b>70.74**</b>	<b>(66.51, 74.97)</b>
RF	50.74	(45.56, 55.92)	<b>62.78**</b>	<b>(51.36, 74.19)</b>	57.78**	(50.93, 64.63)	65.00*	(35.05, 94.95)	54.44**	(49.26, 59.63)
GB	63.33	(48.39, 78.28)	49.44	(46.18, 52.71)	44.44**	(44.44, 44.44)	65.56	(60.12, 70.99)	<b>67.04</b>	<b>(52.09, 81.98)</b>
XGBoost	51.11	(48.21, 54.01)	<b>59.26**</b>	<b>(59.26, 59.26)</b>	56.67**	(54.89, 58.45)	48.15	(48.15, 48.15)	54.81**	(51.91, 57.72)

<sup>3,4</sup>The \* represents p-value < 0.05 and \*\* represents p-value < 0.01 to show the significance of the improved performances over the baseline models. The best results of each learning model are highlighted as bold.

**Table 4**Performance Comparison of Specificity<sup>4</sup>

Models	Structured data only		+ Data-driven CNF		+ Data-driven AF + FS		+ Data-driven and Knowledge-driven CNF		+ Augmented prediction using CNF	
	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)
LR	70.34	(58.71, 81.97)	<b>79.13**</b>	<b>(78.75, 79.51)</b>	80.68**	(80.68, 80.68)	67.05	(51.96, 82.15)	69.86	(58.23, 81.48)
SVM	68.99	(62.33, 75.64)	<b>82.56**</b>	<b>(72.54, 92.58)</b>	79.37**	(69.61, 89.13)	70.24	(56.16, 84.32)	68.50	(61.85, 75.15)
MLP	68.99	(64.78, 73.19)	80.05**	<b>(72.16, 87.94)</b>	<b>87.92**</b>	(85.42, 90.43)	69.81	(66.93, 72.69)	68.50	(64.29, 72.71)
RF	<b>82.42</b>	<b>(77.49, 87.34)</b>	74.69	(63.63, 85.75)	81.45	(74.24, 88.66)	70.63	(41.35, 99.90)	81.93	(77.01, 86.86)
GB	68.74	(53.64, 83.85)	74.64	<b>(70.38, 78.90)</b>	<b>84.54**</b>	(84.54, 84.54)	67.73	(62.70, 72.75)	67.78	(52.67, 82.89)
XGBoost	83.77	(80.15, 87.39)	81.93	(80.80, 83.07)	84.78	(83.57, 86.00)	<b>88.55**</b>	(88.12, 88.98)	83.29	(79.66, 86.91)

<sup>3,4</sup>The \* represents p-value < 0.05 and \*\* represents p-value < 0.01 to show the significance of the improved performances over the baseline models. The best results of each learning model are highlighted as bold.