*JAMIA Open*, 2023, **6(4)**, ooad087
https://doi.org/10.1093/jamiaopen/ooad087
**Research and Applications**

Research and Applications

# Development and application of pharmacological statin-associated muscle symptoms phenotyping algorithms using structured and unstructured electronic health records data

**Boguang Sun, PharmD[1], Pui Ying Yew, BS[2], Chih-Lin Chi, PhD[2,3], Meijia Song, BS[3], Matt Loth, PhD[4], Rui Zhang, PhD[2,4], Robert J. Straka, PharmD*[,1]**

[1]Department of Experimental and Clinical Pharmacology, University of Minnesota College of Pharmacy, Minneapolis, MN 55455, United States, [2]Institute for Health Informatics, Office of Academic Clinical Affairs, University of Minnesota, Minneapolis, MN 55455, United States, [3]School of Nursing, University of Minnesota, Minneapolis, MN 55455, United States, [4]Center for Learning Health System Sciences, University of Minnesota Medical School, Minneapolis, MN 55455, United States

*Corresponding author: Robert J. Straka, PharmD, FCCP, Department of Experimental and Clinical Pharmacology, University of Minnesota College of Pharmacy, 7-115B Weaver-Densford Hall, Minneapolis, MN 55455 (strak001@umn.edu)

## Abstract

**Importance:** Statins are widely prescribed cholesterol-lowering medications in the United States, but their clinical benefits can be diminished by statin-associated muscle symptoms (SAMS), leading to discontinuation.

**Objectives:** In this study, we aimed to develop and validate a pharmacological SAMS clinical phenotyping algorithm using electronic health records (EHRs) data from Minnesota Fairview.

**Materials and Methods:** We retrieved structured and unstructured EHR data of statin users and manually ascertained a gold standard set of SAMS cases and controls using the published SAMS-Clinical Index tool from clinical notes in 200 patients. We developed machine learning algorithms and rule-based algorithms that incorporated various criteria, including ICD codes, statin allergy, creatine kinase elevation, and keyword mentions in clinical notes. We applied the best-performing algorithm to the statin cohort to identify SAMS.

**Results:** We identified 16 889 patients who started statins in the Fairview EHR system from 2010 to 2020. The combined rule-based (CRB) algorithm, which utilized both clinical notes and structured data criteria, achieved similar performance compared to machine learning algorithms with a precision of 0.85, recall of 0.71, and F1 score of 0.77 against the gold standard set. Applying the CRB algorithm to the statin cohort, we identified the pharmacological SAMS prevalence to be 1.9% and selective risk factors which included female gender, coronary artery disease, hypothyroidism, and use of immunosuppressants or fibrates.

**Discussion and Conclusion:** Our study developed and validated a simple pharmacological SAMS phenotyping algorithm that can be used to create SAMS case/control cohort to enable further analysis which can lead to the development of a SAMS risk prediction model.

## Lay Summary

Statins are commonly prescribed cholesterol-lowering medications in the United States, but some patients may experience statin-associated muscle symptoms (SAMS) that can reduce their benefits. In this study, we developed and tested a simple algorithm using electronic health records (EHRs) to identify cases of SAMS. We retrieved data from statin users in the Minnesota Fairview EHR system and manually identified a gold standard set of SAMS cases and controls using a clinical tool. We developed machine learning and rule-based algorithms that considered various criteria, such as ICD codes, statin allergy, creatine kinase elevation, and keyword mentions in clinical notes. The best-performing algorithm, called the combined rule-based (CRB) algorithm, achieved similar performance to machine learning algorithms in identifying SAMS cases. When applied to the larger statin cohort, the CRB algorithm identified a prevalence of 1.9% for pharmacological SAMS and identified selective risk factors such as female gender, coronary artery disease, hypothyroidism, and use of immunosuppressants or fibrates. The developed algorithm has the potential to help create SAMS case/control cohorts for future studies such as building models to predict SAMS risks for patients.

**Key words:** hydroxymethylglutaryl-CoA reductase inhibitors; electronic health records; phenotyping; machine learning; precision medicine.

## Introduction

Nearly half of Americans over 65 years of age take statins, a class of cholesterol-lowering medications proven to reduce morbidity and mortality.[1] However, around 25%–50% of statin users do not fully experience the benefits of statins because of statin discontinuation.[2] Among the reasons for statin discontinuation are personal preference, financial burdens, or side effects. Around 25% of former statin users attributed their non-adherence or discontinuation to side effects, predominantly statin-associated muscle symptoms (SAMS).[2]

**Received:** May 24, 2023; **Revised:** August 3, 2023; **Editorial Decision:** August 9, 2023; **Accepted:** October 3, 2023

Post-market pharmacovigilance of adverse drug reactions (ADRs) including SAMS, is crucial to ensure that medications are safe in the long term when used in real-world settings.[3] FDA Adverse Event Reporting System (FAERS) is one such well-recognized pharmacovigilance program for all approved medications and therapeutic biologics.[4] However, studies have found underreporting of certain ADRs in the FAERS dataset compared to the real-world evidence.[5,6] Furthermore, clinicians might not routinely report certain ADRs to the FDA, especially when they are familiar or insidious, as is often the case with SAMS. Therefore, in order to optimize the appropriate use of statins, there is a critical need to identify the predictors of the development of SAMS based on real-world data where there is sufficient documentation of longitudinal use.

In recent years, with the increasing usage of electronic health records (EHRs) as patient data warehouses, targeted mining of real-world data stored in EHRs has garnered attention as an alternative means for ADR detection and monitoring.[7] Specific phenotyping of ADRs can facilitate the cohort identification and creation for downstream analysis such as genome-wide association studies[3] and risk prediction model development.[4] In EHRs, the minority of patient-centered data is in structured format such as procedures and laboratory tests, whereas the majority takes the form of unstructured data consisting of clinical notes in the free-text format.[7–9] Signals within EHRs that offer evidence for SAMS manifestations include International Classification of Diseases (ICD) coding of muscle symptoms such as myopathy and myalgia, patients' allergy list specific to statin intolerance, temporal creatine kinase (CK) elevation, and most importantly, clinicians' notes documenting the incidence and development of SAMS during patient visits.

To date, SAMS clinical phenotyping algorithms developed based on various EHR systems have shown that a combination of structured and unstructured SAMS-related EHR signals can better identify SAMS compared to using structured data alone.[10,11] However, current studies have not investigated the specific phenotyping of pharmacological SAMS (non-nocebo SAMS). To that end, we aim to develop and validate a pharmacological SAMS clinical phenotyping algorithm based on the University of Minnesota's (UMN) Clinical Data Repository (CDR) with a coverage of Fairview EHRs which includes information from 6 hospitals and over 115 clinics within Minnesota. We applied a scalable Natural Language Processing-Patient Information Extraction for Research (NLP-PIER) tool integrated within the EHR database to search for clinical notes associated with SAMS.[12] We utilized the validated SAMS-Clinical Index (SAMS-CI) tool to ascertain pharmacological SAMS and develop gold standard manual annotations for our phenotyping algorithm.[13] We also applied the best-performing phenotyping algorithm to classify the SAMS (case) and non-SAMS (control) cohorts and reported the differences in patient characteristics and risk factors associated with SAMS. Our phenotyping algorithm can be utilized to differentiate pharmacological SAMS from nocebo SAMS, which was not achieved by medical codings alone. The accurate identification of pharmacological SAMS cases by our phenotyping algorithms can enhance our understanding of the real-world pattern of their occurrence and pharmacologically relevant risk factors, thus enabling construction of pharmacological SAMS risk prediction models using real-world data. Deployment of effective risk prediction models could therefore be incorporated in clinical decision support tools used by clinicians selecting statins for individual patients.

## Methods

### Data source and cohort identification

We retrieved our study cohort from Fairview EHR between January 1, 2010 to December 31, 2020, which represents our study period. As shown in Figure 1, the overall statin cohort contains patients over 18 years old at index date and were regular Fairview system users. We defined regular Fairview system users as having at least one record of each of the following during both the baseline and follow-up periods: (1) Fairview encounter, (2) blood pressure or weight measurements, (3) Fairview pharmacy dispensing, and (4) laboratory data.[14]

Index date was the day the patient was prescribed their first statin medications (atorvastatin, fluvastatin, lovastatin, pitavastatin, pravastatin, rosuvastatin, and simvastatin). The baseline period used to define demographic, comorbidity, and social history was a year preceding the index date. The baseline period to define co-medications was 3 months preceding the index date. The follow-up period was 1 year after the index date or the end of the study period, whichever was earlier.

For the statin cohort, we included patients who initiated any statins and were regular Fairview EHR users during the study period. To exclude prevalent statin users, we excluded any patient who had any statin prescriptions prior to the index date.

We retrieved and analyzed structured EHR data within the relational databases containing patient demographics, medications, laboratory, and procedures maintained by UMN Academic Health Center-Information Exchange (AHC-IE) team. We obtained and searched for clinical notes related to SAMS using the NLP-PIER tool, an NLP search engine enabled by AHC-IE.[12] Figure 2 demonstrates the overview of study workflow and methodology. The study is approved by University of Minnesota IRB (STUDY00011134).

### Manual case ascertainment

To examine the structure and documentation styles of clinical notes within Fairview EHR, we randomly selected clinical notes from 100 patients where their notes included any mentioning of a named statin medication 10 words before or after mentioning reference to any muscle complaints such as muscle pain, myalgia, or myopathy after the index date. Then, we created the NLP-PIER search term that includes mentions of statin medications, muscle symptoms and excludes the negation phrases such as "no myalgia" or "deny myopathy." Next, we created the gold standard set using clinical notes from another independent 200 patients. These 200 patients consisted of a balanced number of potential SAMS cases and non-SAMS controls classified by the NLP-PIER search term.

Two domain experts (with either pharmacy or nursing backgrounds) manually reviewed and annotated the clinical notes in the gold standard set. The manual reviewers annotated and ascertained the pharmacological SAMS cases based on the SAMS-CI tool.[13] This tool aims to discern pharmacological SAMS from nocebo SAMS by incorporating muscle distribution of symptoms, temporal patterns (symptom onset after statin initiation, improvement, and recurrence after statin discontinuation and rechallenge) into a scoring system. Where needed, a third domain expert was consulted to

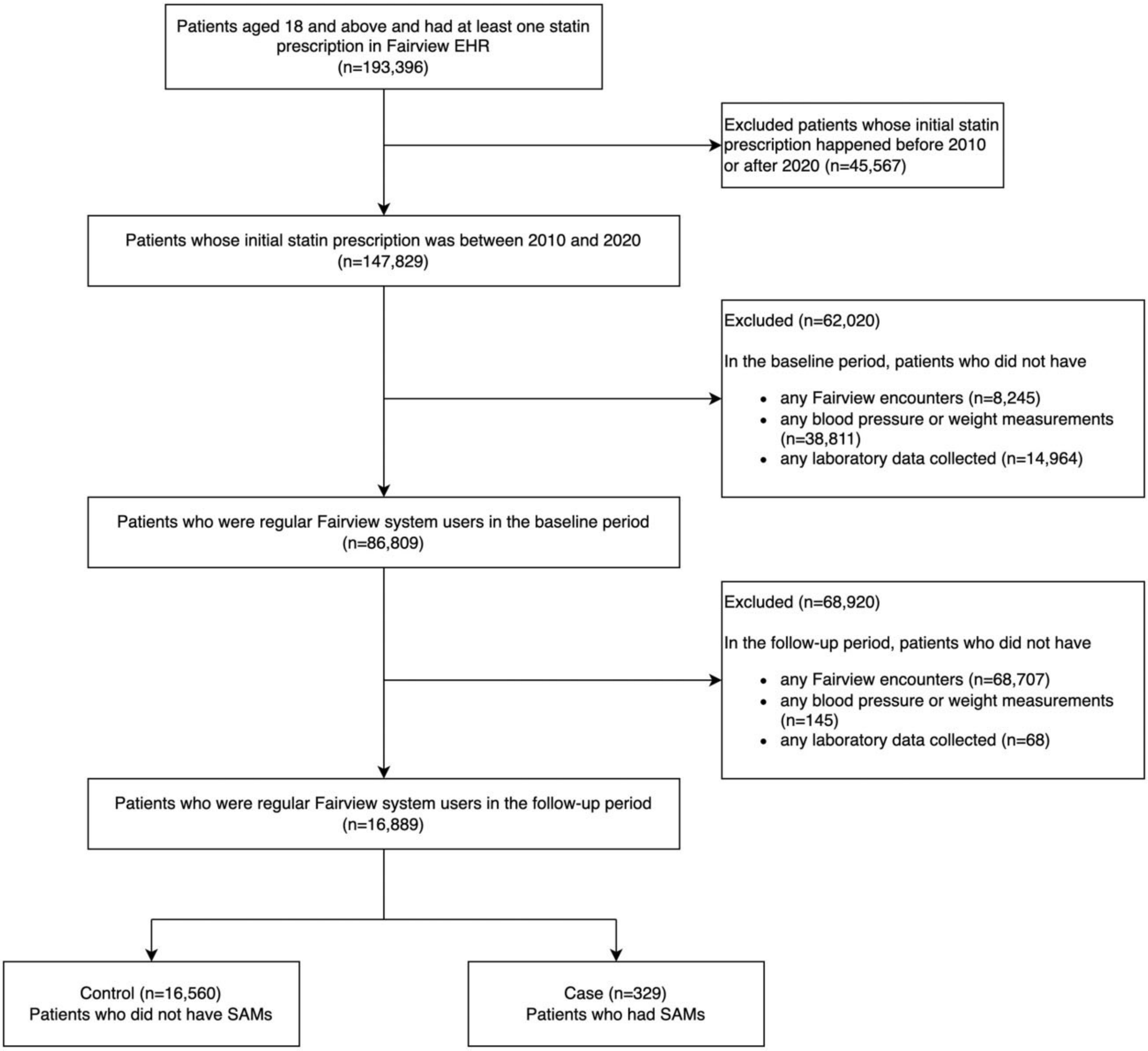

Patients aged 18 and above and had at least one statin prescription in Fairview EHR (n=193,396)
Excluded patients whose initial statin prescription happened before 2010 or after 2020 (n=45,567)
Patients whose initial statin prescription was between 2010 and 2020 (n=147,829)
Excluded (n=62,020)
In the baseline period, patients who did not have
• any Fairview encounters (n=8,245)
• any blood pressure or weight measurements (n=38,811)
• any laboratory data collected (n=14,964)
Patients who were regular Fairview system users in the baseline period (n=86,809)
Excluded (n=68,920)
In the follow-up period, patients who did not have
• any Fairview encounters (n=68,707)
• any blood pressure or weight measurements (n=145)
• any laboratory data collected (n=68)
Patients who were regular Fairview system users in the follow-up period (n=16,889)
Control (n=16,560) Patients who did not have SAMs
Case (n=329) Patients who had SAMs

**Figure 1.** Study flowchart for statin cohort selection from Fairview EHR. Abbreviation: EHR, electronic health record.

reconcile any discrepancies. This tool was also prospectively utilized to ascertain SAMS in an ongoing clinical trial.[15]

Table S1 provides information on how domain experts interpreted the clinical scenario and assigned scores based on notes referencing clinical manifestations of SAMS according to the published SAMS-CI scoring symptoms and clinical scenarios for each score assignment. In Table S2, we provided 2 case vignettes to further illustrate how we determined the scores for each patient.

Overall, we assigned patients with a score greater or equal to 7 points as SAMS cases according to the SAMS-CI tool.[13] We used Cohen's kappa values to assess the manual review agreement between the reviewers in the gold standard set.

## Algorithm development: rule-based and machine learning (ML) algorithms

We considered 6 rule-based algorithms: (1) ICD codes only; (2) allergy list only; (3) CK elevation only; (4) structured components only: ICD codes or allergy list or CK elevation; (5) unstructured component only: clinical notes mentions only; and (6) combination of structured data (ICD codes, allergy list, and CK elevation) and clinical notes mentions. The follow-up period for each individual criterion (ICD codes, allergy list, CK elevation, and clinical notes mentions) was one year after the index date. We selected a 1-year follow-up because this was the timeframe that most statin adverse events occur.[16] To further evaluate the predictive capability of SAMS signals beyond 1 year, we conducted a sensitivity analysis. We explored whether extending the follow-up period at 2 and 4 years would improve the rule-based phenotyping algorithm performance.

*ICD criterion*: Table S3 shows the specific ICD codes we included as signals for SAMS. Patients met the ICD criterion if they only had ICD codes after the index date (no prior ICD codes documentation of muscle symptoms). By implementing this approach, we aimed to exclude patients with pre-existing muscle conditions from being classified as potential SAMS cases in our study.

*Allergy criterion*: Patients met allergy criterion if their allergy list in the EHR indicated having muscle symptoms due to statin medications.

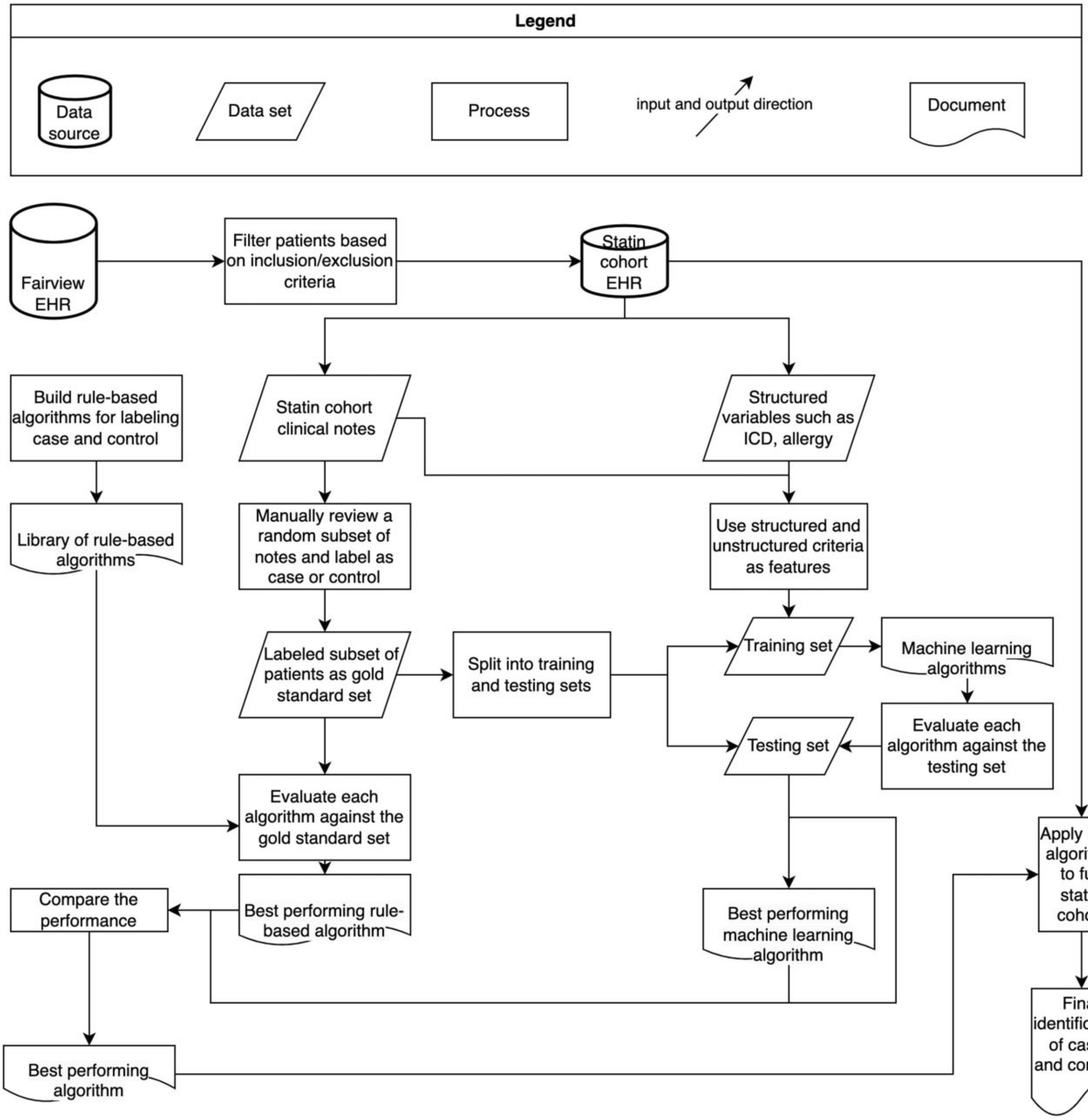

Legend
Data source
Data set
Process
input and output direction
Document
Fairview EHR
Filter patients based on inclusion/exclusion criteria
Statin cohort EHR
Build rule-based algorithms for labeling case and control
Statin cohort clinical notes
Structured variables such as ICD, allergy
Library of rule-based algorithms
Manually review a random subset of notes and label as case or control
Use structured and unstructured criteria as features
Labeled subset of patients as gold standard set
Split into training and testing sets
Training set
Machine learning algorithms
Evaluate each algorithm against the testing set
Testing set
Evaluate each algorithm against the gold standard set
Apply Best algorithm to full statin cohort
Compare the performance
Best performing rule-based algorithm
Best performing machine learning algorithm
Best performing algorithm
Final identification of cases and controls

**Figure 2.** Overview of workflow and methodology/library of rule-based algorithm is a selection of SAMS signals (ICD codes, allergy, CK elevation, clinical notes mentions) alone or in combination with others as rule-based labels.

*CK elevation criterion*: We chose to use a threshold of CK >3 times the upper limit of normal according to the SEARCH (Study of the Effectiveness of Additional Reductions in Cholesterol and Homocysteine) trial.[17] The CK normal ranges used were 30–145 U/L for females and 55–170 U/L males.[18]

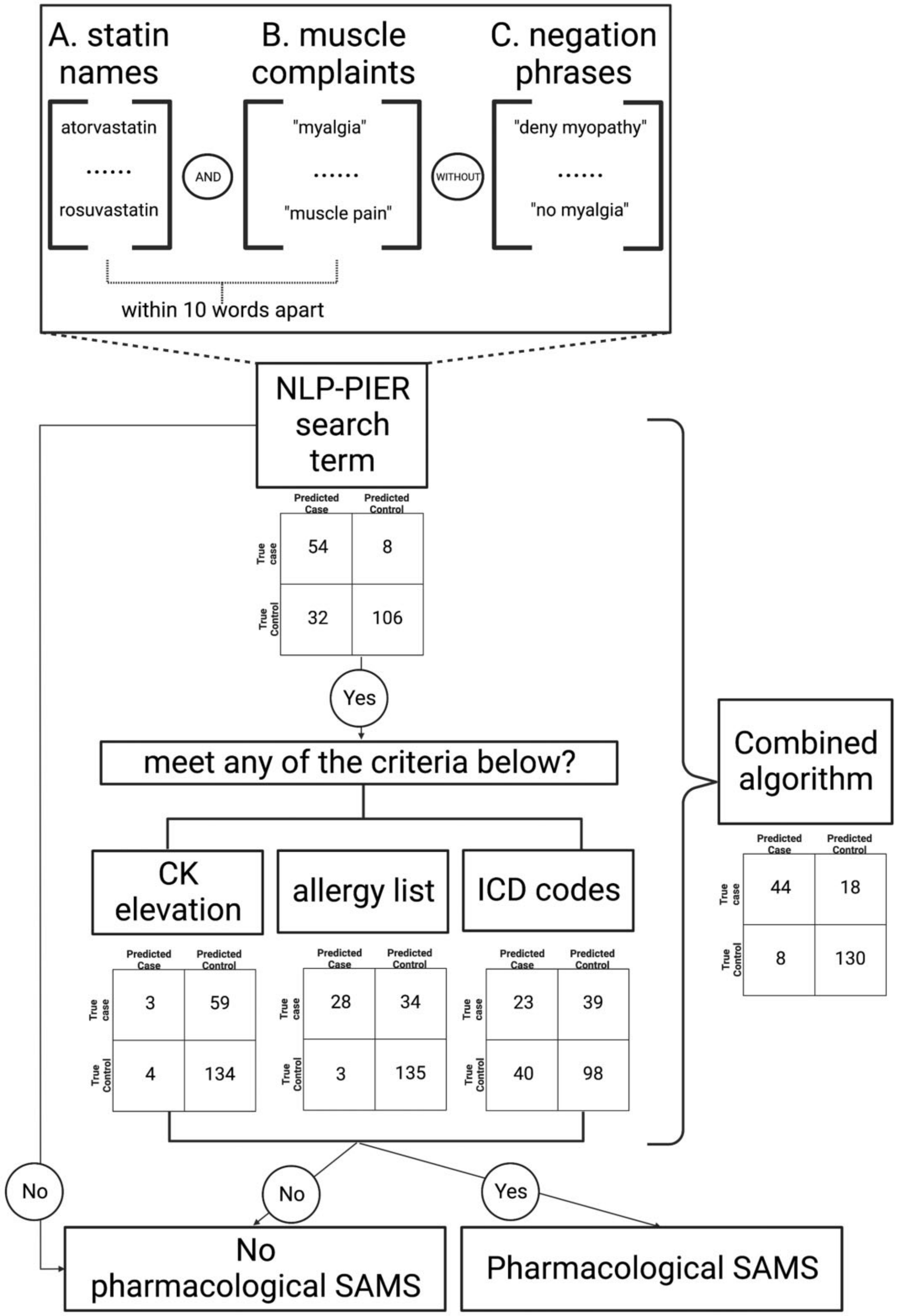
*Clinical notes mentions criterion*: Patients met the notes mentioning criterion if, after the index date, there were any mentions of statin medications 10 words before and after the mentioning of muscle complaints without mentions of negation phrases (NLP-PIER search term).

*Combined rule-based (CRB) algorithm*: Figure 3 shows the decision flowchart for the pharmacological SAMS identification CRB algorithm. Specifically, the CRB algorithm determined the patient to have pharmacological SAMS when they met (1) the clinical notes mentions criterion and; (2) at least one of the structured data criteria (ICD codes, CK elevation, or allergy list). Of note, if a patient had occurrences of a single rule-based signal, only one instance was counted as sufficient to meet the criterion.

For the ML algorithms, we used common ML classifiers including Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and AdaBoost (AB). These algorithms were selected based on their commonly used and well-established nature when applied in similar tasks.[19,20] We split the gold standard set into 70% training and 30% testing set. We used the 4 rule-based labels (ICD codes, allergy list, CK elevation, and clinical notes mentions) as binary features to train the ML classifiers.

## Algorithm evaluation and application

We evaluated the rule-based algorithms against the whole gold standard set. Next, we compared the best-performing rule-based algorithm with ML algorithms against the whole gold standard set (also referred to as the overall set) with 10-



|  | Predicted Case | Predicted Control |
|---|---|---|
| True case | 54 | 8 |
| True Control | 32 | 106 |

|  | Predicted Case | Predicted Control |
|---|---|---|
| True case | 3 | 59 |
| True Control | 4 | 134 |

|  | Predicted Case | Predicted Control |
|---|---|---|
| True case | 28 | 34 |
| True Control | 3 | 135 |

|  | Predicted Case | Predicted Control |
|---|---|---|
| True case | 23 | 39 |
| True Control | 40 | 98 |

|  | Predicted Case | Predicted Control |
|---|---|---|
| True case | 44 | 18 |
| True Control | 8 | 130 |

**Figure 3.** Flowchart for the individual and combined rule-based (CRB) algorithms.

fold cross-validation and the gold standard testing set (referred to as the testing set thereafter). The gold standard testing set is a subset within the whole gold standard set as specified in the previous section. We reported the performance of each algorithm (precision, recall, and F1 scores) in this binary classification problem to separate patients into cases (have pharmacological SAMS) or controls (do not have pharmacological SAMS).

Then, we applied the best-performing algorithm to the overall statin cohort. We reported preliminary patient baseline characteristics between cases and controls including demographics, social history, comorbidities, and concurrent medications associated with pharmacological SAMS risk.[21] We conducted univariate and multivariate association analysis of pharmacological SAMS outcome and risk factors. We included statistically significant ($P < 0.05$)

baseline factors from the univariate analysis in the multivariate analysis.

## Results

### Cohort identification

As shown in Figure 1, out of the 193 396 adult patients who started statins in the Fairview EHR system, we included 16 889 patients who met our criteria in this study: patients who started statins during 2010 to 2020, were regular Fairview EHR users, and were not prevalent statin users.

### Manual case ascertainment

Two reviewers annotated clinical notes from 200 patients in the gold standard set. The 2 reviewers achieved high agreement in determining the case vs controls using the SAMS-CI tool (kappa = 0.985).

In the gold standard set, the NLP-PIER search term identified 86 cases and 114 controls. After manual reviews, we ascertained 62 cases and 138 controls (true cases and true controls).

### Algorithm development and evaluation

As illustrated in Figure 3, the NLP-PIER search term for SAMS has 3 components: (A) mentioning of any statin medications (see Methods for the statin medications list); (B) mentioning of any muscle complains including "myalgia," "myopathy," "muscle pain," "muscle ache," "muscle cramp," "myositis"; and (C) with the mentioning of negation phrases including "no myalgia," "no myopathy," "deny myalgia," "deny myopathy," "suspect myalgia," "no muscle aches," "monitor for myalgia." The mentioning of criteria A and B has to be within 10 words apart. In Figure 2, we also reported the algorithm performance in individual and CRB algorithms using the confusion matrices from the gold standard set (200 patients).

As shown in Table 1, the CRB algorithm achieved better performances compared to the other rule-based algorithms. The precision, recall, and F1 score were 0.85, 0.71, 0.77 against the gold standard set ($N = 200$), respectively. The ICD only algorithm had the worst performance compared with other algorithms with an F1 score of only 0.37 against the gold standard set. The allergy only algorithm had good performance in terms of precision (0.90 against the gold standard set) but its recall was compromised (0.45 against the gold standard set). The notes-only algorithm had better performance in terms of recall compared to the CRB algorithm but was outperformed by the CRB algorithm regarding the precision (0.63 against the gold standard set).

As shown in Table 2, the CRB rule-based algorithm achieved similar performances compared to the ML algorithms in the overall set ($N = 200$) with 10-fold cross-validation and testing set ($N = 60$). The RF classifier had slightly better recall than the CRB algorithm when evaluating against the overall set but the differences in recall were diminished when compared against the testing set.

As shown in Table S4, the sensitivity analysis demonstrated that utilizing the SAMS rule-based labels beyond 1 year for phenotyping did not achieve similar performances as compared to defining them at 1 year time frame.

**Table 1.** Rule-based algorithm performances by precision, recall, and F1 scores.

| | Precision | Recall | F1 score |
|---|---|---|---|
| ICD only | 0.37 | 0.37 | 0.37 |
| Allergy only | 0.90 | 0.45 | 0.60 |
| CK elevation only | 0.43 | 0.05 | 0.09 |
| ICD or allergy or CK elevation | 0.53 | 0.77 | 0.63 |
| Notes only | 0.63 | 0.87 | 0.73 |
| Combined rule-based algorithm | 0.85 | 0.71 | 0.77 |

**Table 2.** Combined rule-based and machine learning algorithm performances.

| Algorithms | Precision | Recall | F1 score |
|---|---|---|---|
| Combined rule-based | | | |
| Overall set ($N = 200$) | 0.85 | 0.71 | 0.77 |
| Testing set ($N = 60$) | 0.89 | 0.84 | 0.86 |
| Random forest | | | |
| Overall set[a] | 0.63 | 0.80 | 0.70 |
| Testing set | 0.80 | 0.84 | 0.82 |
| Adaptive boosting | | | |
| Overall set | 0.82 | 0.67 | 0.73 |
| Testing set | 0.84 | 0.84 | 0.84 |
| K-nearest neighbors | | | |
| Overall set | 0.86 | 0.68 | 0.74 |
| Testing set | 0.88 | 0.79 | 0.83 |
| Decision tree | | | |
| Overall set | 0.89 | 0.55 | 0.68 |
| Testing set | 0.82 | 0.47 | 0.60 |

[a] Overall set performances (precision, recall, and F1 score) were mean values from 10-fold cross-validation.

### Algorithm applications: patient characteristics comparison and association analysis

After applying the CRB algorithm to the statin cohort, we identified 329 cases and 16 560 controls. This translated to a pharmacological SAMS prevalence of 1.9% (329/16 889) in our statin cohort. Table 3 shows the baseline characteristics of the cases and controls. Briefly, the mean age was 67.1 vs 66.8 in cases and controls, respectively. The pharmacological SAMS case group had significantly more females than the controls (50.5% vs 44.5%, $P < 0.05$). Additionally, the SAMS cases group had significantly more hypertension (74.2% vs 66.3%), coronary artery disease (52.9% vs 37.3%), chronic kidney disease (11.9% vs 7.6%), and hypothyroidism (20.1% vs 12.8%) than the controls. Significantly more patients in the cases group took beta-blockers (53.5% vs 45%), immunosuppressants (13.7% vs 8.3%), and fibrates (4.5% vs 2.2%) than the controls.

As shown in Table 4, all the baseline factors shown to be significantly different between cases and controls were also significant risk factors identified using univariate analysis. However, only female gender, coronary artery disease, hypothyroidism, and use of immunosuppressant and fibrates were associated with higher risk of SAMS after the multivariate analysis.

## Discussion

Studies[22,23] have demonstrated that EHRs can be used as a reliable source for ADR phenotyping and downstream

**Table 3.** Demographic and baseline characteristics of the pharmacological SAMS case and control patients.[a]

| | Case (*N* = 329) | Control (*N* = 16 560) | *P* values |
|---|---|---|---|
| Age (years) | 67.1 ± 12.8 | 66.8 ± 13.9 | 0.58 |
| Female sex | 166 (50.5) | 7374 (44.5) | **0.04**[b] |
| Race[c] | | | |
| White | 296 (89.9) | 14 489 (87.5) | 0.6 |
| Asian | 7 (2.1) | 385 (2.3) | |
| Black | 15 (4.6) | 854 (5.2) | |
| Other | 11 (3.4) | 832 (5) | |
| Body mass index (kg/m$^2$) | 30.1 ± 6.5 | 29.9 ± 7.6 | 0.55 |
| Social history | | | |
| Smoking | 98 (29.8) | 5403 (32.6) | 0.3 |
| Alcohol | 150 (45.6) | 7121 (43) | 0.38 |
| Medical history | | | |
| Hypertension | 244 (74.2) | 10985 (66.3) | **<0.01** |
| Diabetes mellitus | 105 (31.9) | 5121 (30.9) | 0.75 |
| Coronary artery disease | 174 (52.9) | 6176 (37.3) | **<0.01** |
| Congestive heart failure | 78 (23.4) | 3360 (20.3) | 0.15 |
| Chronic kidney disease | 39 (11.9) | 1257 (7.6) | **<0.01** |
| Hypothyroidism | 68 (20.1) | 2122 (12.8) | **<0.01** |
| Medication history | | | |
| Angiotensin-converting enzyme inhibitors | 101 (30.1) | 4579 (27.7) | 0.25 |
| Beta-blockers | 176 (53.5) | 7458 (45) | **<0.01** |
| Immunosuppressants[d] | 45 (13.7) | 1367 (8.3) | **<0.01** |
| Fibrates[e] | 14 (4.5) | 370 (2.2) | **0.02** |

[a] Baseline characteristics were defined within one year preceding the index date.
[b] Bolding denotes statistical significance.
[c] Categorical variables in count (%), while continuous variables in mean (±standard deviation).
[d] Immunosuppresants include cyclosporine, everolimus, sirolimus, tacrolimus.
[e] Fibrates include fenofibrates and gemfibrozil.

**Table 4.** Univariate and multivariate logistic regression of risk factors and pharmacological SAMS outcome

| | Univariate | | Multivariate | |
|---|---|---|---|---|
| | OR (95% CI) | *P* values | OR (95% CI) | *P* values |
| Age | 1 (0.99, 1.01) | 0.6 | | |
| Female sex | 1.27 (1.02, 1.59) | 0.033 | 1.33 (1.05, 1.67) | 0.02 |
| Body mass index | 0.89 (0.82, 1.21) | 0.46 | | |
| Smoking | 0.88 (0.69, 1.11) | 0.3 | | |
| Alcohol use | 1.11 (0.89, 1.38) | 0.3 | | |
| Hypertension | 1.46 (1.14, 1.88) | 0.003 | 1.26 (0.98, 1.63) | 0.08 |
| Diabetes mellitus | 1.05 (0.83, 1.32) | 0.7 | | |
| Coronary artery disease | 1.89 (1.52, 2.35) | <0.001 | 1.84 (1.47, 2.32) | <0.001 |
| Congestive heart failure | 1.22 (0.94, 1.57) | 0.13 | | |
| Chronic kidney disease | 1.64 (1.15, 2.27) | 0.004 | 1.28 (0.88, 1.79) | 0.18 |
| Hypothyroidism | 1.77 (1.34, 2.31) | <0.001 | 1.59 (1.19, 2.09) | <0.001 |
| Angiotensin-converting enzyme inhibitors | 1.1 (0.98, 1.23) | 0.62 | | |
| Beta-blockers | 1.4 (1.13, 1.75) | 0.002 | 1.16 (0.92, 1.46) | 0.2 |
| Immunosuppressants | 1.76 (1.26, 2.4) | <0.001 | 1.66 (1.18, 2.28) | <0.001 |
| Fibrates | 1.94 (1.08, 3.23) | 0.017 | 1.93 (1.07, 3.22) | 0.02 |

Abbreviations: CI, confidence interval; OR, odds ratio.

research such as pharmacovigilance and genetics studies.[24,25] SAMS, as an example of ADR, has been challenging for phenotyping due to heterogeneity of symptoms and nocebo effects.[26] As a result, the prevalence of all SAMS ranges from 1% to 25%, but the prevalence of pharmacological (non-nocebo) SAMS is estimated to be only about 1%–2%.[20] To date, multiple studies have proposed SAMS phenotyping algorithms. Duke et al used ICD-9 codes to define myopathy and rhabdomyolysis.[27] Other studies investigated the usage of CK elevation to define SAMS or statin myopathy.[27,28] Additionally, NLP-enabled approaches that combine structured and unstructured data components to define SAMS have been successful.[9,10] Specifically, Willey et al have developed a statin myopathy keyword filter that achieved superior performance compared to using structured data alone such as ICD codes and CK elevation.[10]

However, these above-mentioned studies[10,11,27,28] have utilized different case ascertainment methods without the use of the SAMS-CI tool, which is specifically designed to differentiate pharmacological SAMS from nocebo SAMS. Furthermore, the above phenotyping algorithms[10,11,27,28] have not been scaled and applied to create case/control datasets for downstream applications. Given these gaps in the literature, our study was motivated by the need to develop a pharmacological SAMS phenotyping algorithm using EHR data from the Fairview Healthcare system. We aimed to leverage the available EHR data, including structured data and unstructured clinical notes, to develop a robust algorithm for identifying pharmacological SAMS. Subsequently, we applied this algorithm to create a case and control cohort for a future development of pharmacological SAMS risk prediction models.

In this study, we first identified the statin user cohort and defined statin index date, baseline, and follow-up periods. These timelines were crucial for us to analyze the temporal relationship between statin use and muscle symptoms and calculate the SAMS-CI score. We also defined regular Fairview EHR users to ensure that the patients included had sufficient longitudinal clinical notes of their system encounters. Specifically, each patient in our cohort had ∼40 statin-related clinical notes. This allowed us to leverage more information within the clinical notes to sufficiently adjudicate the SAMS cases vs controls using the SAMS-CI tool.

We developed pharmacological SAMS phenotype algorithms using structured and unstructured EHR data in an integrated healthcare system. As demonstrated in Table 1, using structured data components alone or in combination such as ICD coding, allergy list, or CK elevation as phenotyping algorithms could not identify pharmacological SAMS with reasonable performance. Using clinical notes mentions as a single criterion for SAMS can achieve similar recalls compared to the CRB algorithm but it did not perform well in terms of precision (high false positive rates). Overall, the CRB algorithm with consideration of patients' allergy list, ICD coding of muscle symptoms, CK elevation, and clinical notes mentions achieved the best performance for pharmacological SAMS identification. We designed the CRB algorithm in a hierarchical structure where we gave the clinical notes mentions criterion more weight in determining the cases but also leveraged the other criteria to help increase the performance. Of note, our hierarchical CRB algorithm had overall similar performances when compared with ML algorithms (Table 2). The ML algorithms such as RF had incremental improvement in recall compared to the CRB algorithm. However, since our end-goal was to use the best-performing phenotyping algorithm to classify SAMS cases and controls, high precision becomes a more desirable metric in our model evaluation. Additionally, the rule-based algorithm also has clinical advantage as it is easier to interpret. Overall, we chose the CRB algorithm as the best-performing algorithm for application.

Our study applied the CRB algorithm on the pre-defined statin cohort ($N = 16\,889$) as shown in Figure 2. We estimated the prevalence of pharmacological SAMS to be 1.9% (329/16 889), which was similar to the estimation reported in the current National Lipid Association guidelines.[21] As shown in Table 3, the prevalence or values of several baseline factors were statistically different between SAMS case ($N = 329$) and control ($N = 16\,560$) cohorts. After a univariate/multivariate analysis shown in Table 4, we recognized several key risk factors such as female gender, coronary artery disease, hypothyroidism, and use of fibrates and immunosuppressants that were associated with increased risk of pharmacological SAMS in our statin cohort. These risk factors identified in our analysis align with common SAMS risk factors in real-world settings[21] and also previously recognized in national guidelines[29] thus further strengthening the potential clinical usability of our phenotyping algorithm.

Our study had some limitations. First of all, in this study, we focused primarily on rule-based and ML algorithms that utilize EHR components (ICD codes, CK elevation, allergy list, and clinical notes mentions) for prediction. We appreciate that novel ML and deep learning NLP approaches leveraging clinical notes might achieve better performances compared to conventional phenotyping algorithms.[30] Therefore, future studies are needed to develop pharmacological SAMS phenotyping algorithms using novel NLP techniques. Secondly, we acknowledge the potential bias when deploying our algorithm's application in other EHR systems due to the variabilities in population demographics, coding standards, and documentation styles for specific SAMS keywords and language patterns within EHRs. Different institutions may utilize different terminologies or variations in describing SAMS, which could impact the performance of our algorithm in capturing relevant cases. However, we believe that our phenotyping algorithm development framework (Figures 2 and 3) might have the potential for interoperability among different EHR systems. This is because each individual component in the CRB algorithm is readily available in other EHR systems. We also did not over-train the NLP-PIER search term by adding additional filter words. We intended to make the NLP-PIER search term a "weak learner" and when combined with other features such as ICD codes, CK elevation, and allergy list, the model performance was optimized. By providing the basic structure of the search term, institutions can adapt it according to their own requirements.

For future steps, we will develop and validate a pharmacological SAMS risk prediction model using the pharmacological SAMS cases and control cohorts classified by our pharmacological SAMS phenotyping algorithm. We envision that the risk prediction model can be incorporated into patients' EHRs as an element of clinical decision support. Once a patient has an indication for a statin and at the same time been deemed as high risk for developing SAMS, a "warning or cautionary message" could fire in the EHR, prompting a review by prescribers so that preemptive measures (adjustment of doses and selection of specific statin, reviews of interacting medications, and more frequent monitoring, etc.) can be taken to improve statin adherence.

## Conclusion

In this study, we developed a pharmacological SAMS phenotyping algorithm using structured and unstructured data within the Fairview EHRs. The CRB algorithm incorporating unstructured and structured data outperformed all other rule-based algorithms with precision, recall, and F1 of 0.85, 0.71, 0.77 against the gold standard set, respectively. The CRB algorithm also had comparable performances to ML algorithms. We applied the best-performing CRB algorithm on the statin cohort and identified the pharmacological SAMS prevalence of 1.9% and pharmacological SAMS risk factors including female gender, coronary artery disease, hypothyroidism, and use of immunosuppressants and fibrates. These observations align with the real-world clinical practice estimates of pharmacological SAMS which further corroborate the clinical utility of our algorithm.

## Author contributions

B.S. conceived and designed the study, conducted data analysis, interpreted the results, and drafted and revised the article. P.Y.Y., C.-L.C., M.S., and R.Z. contributed to study design, data acquisition, results interpretation, and article revision. R.J.S., C.-L.C., and R.Z. reviewed and further contributed to the article. All authors drafted and revised the article.

## Supplementary material

Supplementary material is available at *JAMIA Open* online.

## Funding

This research was supported by the National Institutes of Health's National Center for Advancing Translational Sciences (grant UL1TR002494). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health's National Center for Advancing Translational Sciences.

## Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Data availability

The proposed combined rule-based algorithm can be reconstructed using the approach as outlined in Figure 3 and supporting text descriptions of criteria as described throughout this article (see Methods, Results, and Supplementary Materials).

The data underlying this article contain Protected Health Information and therefore are securely stored and saved within MHealth Fairview EHR data shelter as required by institutional policy. For investigators seeking access to these data and with the appropriate proof of training, we would be willing to collaborate on requests to add them to our IRB approval to enable access to the data.

## References

1. Horodinschi R-N, Stanescu AMA, Bratu OG, *et al.* Treatment with statins in elderly patients. *Medicina*. 2019;55(11):721.
2. Ingersgaard MV, Helms Andersen T, Norgaard O, Grabowski D, Olesen K. Reasons for nonadherence to statins—a systematic review of reviews. *Patient Prefer Adherence*. 2020;14:675-691.
3. Lavertu A, Vora B, Giacomini KM, Altman R, Rensi S. A new era in pharmacovigilance: toward real-world data and digital monitoring. *Clin Pharmacol Ther*. 2021;109(5):1197-1202.
4. Harpaz R, DuMouchel W, LePendu P, *et al.* Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther*. 2013;93(6):539-546.
5. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opin Drug Saf*. 2017;16(7):761-767.
6. Summers RW, Flatt AJ. A comparative study of the effects of four motor-stimulating agents on canine jejunal spike bursts. The use of a computer program to analyze spike burst spread. *Scand J Gastroenterol*. 1988;23(10):1173-1181.
7. Luo Y, Thompson WK, Herr TM, *et al.* Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf*. 2017;40(11):1075-1089.
8. Kong H-J. Managing unstructured big data in healthcare system. *Healthc Inform Res*. 2019;25(1):1-2.
9. Chiu C-C, Wu C-M, Chien T-N, Kao L-J, Li C, Chu C-M. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent Dirichlet allocation method. *Int J Environ Res Public Health*. 2023;20(5):4340.
10. Wiley LK, Moretz JD, Denny JC, Peterson JF, Bush WS. Phenotyping adverse drug reactions: statin-related myotoxicity. *AMIA Jt Summits Transl Sci Proc*. 2015;2015:466-470.
11. Chan SL, Tham MY, Tan SH, *et al.* Development and validation of algorithms for the detection of statin myopathy signals from electronic medical records. *Clin Pharmacol Ther*. 2017;101(5):667-674.
12. McEwan R, Melton GB, Knoll BC, *et al.* NLP-PIER: a scalable natural language processing, indexing, and searching architecture for clinical notes. *AMIA Jt Summits Transl Sci Proc*. 2016:150-159.
13. Rosenson RS, Miller K, Bayliss M, *et al.* The statin-associated muscle symptom clinical index (SAMS-CI): revision for clinical use, content validation, and inter-rater reliability. *Cardiovasc Drugs Ther*. 2017;31(2):179-186.
14. Mansi IA, Chansard M, Lingvay I, *et al.* Association of statin therapy initiation with diabetes progression: a retrospective matched-cohort study. *JAMA Intern Med*. 2021;181(12):1562-1574.
15. Cha J-J, Hong SJ, Kim JH, *et al.* Effect of rosuvastatin 20 mg versus rosuvastatin 5 mg plus ezetimibe on statin side-effects in elderly patients with atherosclerotic cardiovascular disease: rationale and design of a randomized, controlled SaveSAMS trial. *Am Heart J*. 2023;261:45-50.
16. Cohen JD, Brinton EA, Ito MK, Jacobson TA. Understanding statin use in America and gaps in patient education (USAGE): an internet-based survey of 10,138 current and former statin users. *J Clin Lipidol*. 2012;6(3):208-215.
17. SEARCH Study Collaborative Group; Bowman L, Armitage J, Bulbulia R, Parish S, Collins R. Study of the effectiveness of additional reductions in cholesterol and homocysteine (SEARCH): characteristics of a randomized trial among 12064 myocardial infarction survivors. *Am Heart J*. 2007;154:815-823, 823.e1–6.
18. Pagana KD, Pagana TJ, Pagana TN. *Mosby's Diagnostic and Laboratory Test Reference—E-Book*. Elsevier Health Sciences; 2018.
19. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodical review. *J Am Med Inform Assoc*. 2023;30(2):367-381.
20. Jeong E, Park N, Choi Y, Park RW, Yoon D. Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. *PLoS One*. 2018;13(11):e0207749.
21. Warden BA, Guyton JR, Kovacs AC, *et al.* Assessment and management of statin-associated muscle symptoms (SAMS): a clinical perspective from the national lipid association. *J Clin Lipidol*. 2023;17(1):19-39.
22. Gottesman O, Kuivaniemi H, Tromp G, *et al.*; eMERGE Network. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med*. 2013;15(10):761-771.
23. Chute CG, Pathak J, Savova GK, *et al.* The SHARPn project on secondary use of electronic medical record data: progress, plans, and possibilities. *AMIA Annu Symp Proc*. 2011;2011:248-256.
24. Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3(79):79re1.
25. Dubberke ER, Nyazee HA, Yokoe DS, *et al.* Implementing automated surveillance for tracking Clostridium difficile infection at multiple healthcare facilities. *Infect Control Hosp Epidemiol*. 2012;33(3):305-308.

26. Tobert JA, Newman CB. The nocebo effect in the context of statin intolerance. *J Clin Lipidol*. 2016;10(4):739-747.
27. Duke JD, Han X, Wang Z, *et al.* Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Comput. Biol*. 2012;8(8):e1002614.
28. Sai K, Hanatani T, Azuma Y, *et al.* Development of a detection algorithm for statin-induced myopathy using electronic medical records. *J Clin Pharm Ther*. 2013;38(3):230-235.
29. Grundy SM, Stone NJ, Bailey AL, *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. *Circulation*. 2019;139:e1082-e1143.
30. Zhang R, Ma S, Shanahan L, *et al.* Discovering and identifying New York heart association classification from electronic health records. *BMC Med Inform Decis Mak*. 2018;18:48.