Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm

Chengyin Ye[a], Jinmei Li[a], Shiying Hao[b,c], Modi Liu[d], Hua Jin[d], Le Zheng[b,c], Minjie Xia[d], Bo Jin[d], Chunqing Zhu[d], Shaun T. Alfreds[e], Frank Stearns[d], Laura Kanov[d], Karl G. Sylvester[f], Eric Widen[d], Doff McElhinney[b,c], Xuefeng Bruce Ling[c,f,*]

[a] *Department of Health Management, Hangzhou Normal University, Hangzhou, China*
[b] *Department of Cardiothoracic Surgery, Stanford University, Stanford, CA, United States*
[c] *Clinical and Translational Research Program, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Palo Alto, CA, United States*
[d] *HBI Solutions Inc., Palo Alto, CA, United States*
[e] *HealthInfoNet, Portland, ME, United States*
[f] *Department of Surgery, Stanford University, Stanford, CA, United States*

ARTICLE INFO

Keywords:
Accidental falls
Aged
Electronic health records
Supervised machine learning

ABSTRACT

Objective: Predicting the risk of falls in advance can benefit the quality of care and potentially reduce mortality and morbidity in the older population. The aim of this study was to construct and validate an electronic health record-based fall risk predictive tool to identify elders at a higher risk of falls.

Methods: The one-year fall prediction model was developed using the machine-learning-based algorithm, XGBoost, and tested on an independent validation cohort. The data were collected from electronic health records (EHR) of Maine from 2016 to 2018, comprising 265,225 older patients (≥65 years of age).

Results: This model attained a validated C-statistic of 0.807, where 50 % of the identified high-risk true positives were confirmed to fall during the first 94 days of next year. The model also captured in advance 58.01 % and 54.93 % of falls that happened within the first 30 and 30–60 days of next year. The identified high-risk patients of fall showed conditions of severe disease comorbidities, an enrichment of fall-increasing cardiovascular and mental medication prescriptions and increased historical clinical utilization, revealing the complexity of the underlying fall etiology. The XGBoost algorithm captured 157 impactful predictors into the final predictive model, where cognitive disorders, abnormalities of gait and balance, Parkinson's disease, fall history and osteoporosis were identified as the top-5 strongest predictors of the future fall event.

Conclusions: By using the EHR data, this risk assessment tool attained an improved discriminative ability and can be immediately deployed in the health system to provide automatic early warnings to older adults with increased fall risk and identify their personalized risk factors to facilitate customized fall interventions.

1. Introduction

In adults over 65 years old, fall is recognized as a major cause of injury and hospital admission for trauma and related death worldwide [1]. The estimated incidence rates of fall range from 28 % to 35 % per year for community-dwelling older people, while about 20 % of those fallen people require medical attention, and 5 % would experience fractures and severe head injuries [2]. In the U.S., the total fall-related medical costs were more than $50 billion in 2015 [3], and are expected to reach $55 billion dollars by 2020 [4]. Therefore, providing an early warning tool of fall risk in the older population could alert care-givers of individuals' risk of fall, activate individualized interventions for high-risk individuals, and eventually reduce fall rate and corresponding medical costs.

In the older population, the fall risk factors can be divided into two categories, that is, intrinsic and environmental factors. For intrinsic factors, besides advanced age and gender [4], many disease conditions and physical dysfunctions were reported to associate with an increased

[*] Corresponding author at: Department of Surgery, Stanford University, S370 Grant Bldg., Stanford, CA, United States.

E-mail addresses: yechengyin@hznu.edu.cn (C. Ye), lijinmei@stu.hznu.edu.cn (J. Li), shiyingh@stanford.edu (S. Hao), mliu@hbisolutions.com (M. Liu), hjin@hbisolutions.cn (H. Jin), zhengl07@stanford.edu (L. Zheng), cxia@hbisolutions.com (M. Xia), ejin@hbisolutions.com (B. Jin), czhu@hbisolutions.com (C. Zhu), salfreds@hinfonet.org (S.T. Alfreds), fstearns@hbisolutions.com (F. Stearns), lkanov@hbisolutions.com (L. Kanov), karls@stanford.edu (K.G. Sylvester), Ewiden@hbisolutions.com (E. Widen), Doff@stanford.edu (D. McElhinney), bxling@stanford.edu (X.B. Ling).

https://doi.org/10.1016/j.ijmedinf.2020.104105
Received 30 December 2019; Received in revised form 15 February 2020; Accepted 27 February 2020

risk of fall, such as, muscle weakness, gait and balance problems, poor vision, postural hypotension and many chronic diseases (i.e., osteoporosis, stroke, cognitive impairment, epilepsy and dementia). Furthermore, medications used to treat mental disorders, diabetes and cardiovascular diseases, as well as nonsteroidal anti-inflammatory drugs (NSAID) are also recognized to have strong association with increased fall risk. In terms of extrinsic risk factors, unsafe residence and neighborhood environment are major fall causal effects [5].

Traditional fall risk assessment tools are physical function evaluations that monitor an individual's static and dynamic gait and balance performance [4,6], such as the Timed Up and Go (TUG) test [7]. Another class of assessment tools are generated from fall-related risk factors collected from literatures or questionnaires [8], such as the fall risk model, FRAT-up. However, such meta-analysis-based approach may be subject to low accuracy when population heterogeneity and bias in odds-ratio estimates exist [9]. Since the majority of intrinsic risk factors are enriched in the routinely collected real-time electronic health records (EHR), an EHR-based fall risk assessment tool should be considered as an efficient solution to discover discriminant clinical patterns and crucial triggers of fall. Recently, several models were developed by using patients' outpatient, emergency or inpatient records [10–13], most of which adopted traditional statistical approaches of survival model or cox regressions [11,14]. A multivariate logistic regression model incorporated patients' demographic characteristics, health status, medication and vital signs information to predict unintentional fall risk and attained a retrospective C-statistic of 0.79 [13]. These prediction models' discriminative ability has been significantly improved, especially when the EHR data were updated in high frequency. Nevertheless, the high dimensional EHR data usually requires the integrated algorithms to have the capacity of parallel analysis of thousands of clinical parameters simultaneously, as well as the ability of efficient dimensionality reduction. We wonder whether the accuracy of such tools could be further improved if advanced machine learning algorithms were introduced.

In this study, we aimed to develop an EHR-based risk assessment model to forecast patients' fall risk in the following one year. By using the EHR data from the older population in the State of Maine, U.S., and the advanced machine learning algorithm, we expected that the new model could uncover the underlying clinical and pathophysiological patterns/interactions of impactful predictors, and eventually reach an improved accuracy.

2. Method

2.1. Dataset

The study cohort was formed by patients with age of 65 years and older that visited Maine health care facilities, including 35 hospitals, 34 federally qualified health centers, from April 1, 2016 to March 30, 2018. This retrospective dataset was a subset of the health information exchange (HIE) network and was authorized by the HealthInfoNet organization after the de-identification process. The personal information was removed during the analysis and publication procedure. This study was exempted from ethics review by the Stanford University institutional review board. The inclusion and exclusion criteria were summarized in Fig. 1. A total of 265,225 individuals were recruited in this study.

2.2. Definition of fall and predictive variables

A fall record was defined according to the codes of W00-W119 and R29.6 from the International Classification of Diseases-10 (ICD-10) [1]. To predict the risk of new-incident fall in the following one year, we compiled EMR datasets of the patients for their fall records from April 1, 2017 to March 30, 2018. Patients who suffered multiple falls during the targeted time frame were chart-reviewed by internal physician curators such that only the first fall records were utilized in our analysis. As a result, a total of 4361 falls were identified, and a binary outcome label (1 or 0) was assigned to the cases and controls as the predictive dependent variables.

Accordingly, the candidate predictors were extracted from the EHR dataset during the time period of April 1, 2016 to March 30, 2017, which were mainly demographic characteristics, clinical utilization features, disease diagnosis from ICD-10 codes, medication prescriptions from National Drug Code (NDC), and laboratory test results from Logical Observation Identifiers Names and Codes (LOINC). We sampled the case or control group to randomly divide the same group subjects to 2:1 (training:testing) subgroups. Subsequently, these 2/3 or 1/3 of the case and control subjects were combined to form the training or testing cohort respectively. Therefore, the training or testing dataset was stratified and divided, rather than constructed by pure random.

2.3. Predictive model development and evaluation

Before the modeling process, an initial univariate logistic regression was introduced in the derivation cohort to perform the feature prescreening routine. As a result, a total of 10,198 predictive variables from the EHR dataset passed the significant criteria (p value < 0.05) and were treated as the inputs of the following machine learning process. The predictive model was built on the derivation cohort by two steps. First, several advanced linear or non-linear machine learning algorithms were adopted to construct the predictive model, including Random Forest (RF) [15], XGBoost [16], Lasso [17], K-nearest neighbors (KNN) [18] and Support Vector Machine (SVM) [19]. R libraries of randomForest, xgboost, glmnet, FNN and e1071 were applied respectively. The details of parameter tuning process were carefully introduced for these machine learning algorithms in the Appendix B. Supplementary Methods.

Second, under each derived predictive model, we applied a calibration process based on the positive predictive values (PPVs) to assign an estimated risk score to each individual. The Isotonic Regression, one of the two classical calibration algorithms and suitable for tree-based models [20], was used to do the calibration. To correct any monotonic distortion, Isotonic Regression tried to align the final risk score with the real probability or PPV in a risk bin. In our case, after the initial risk estimates were derived from the above models, a PPV for a certain estimate was calculated as the proportion of cases in the group of patients having risk estimates the same as or larger than this estimate. This PPV, as the measured probability of fall among patients receiving initial risk estimates the same as or larger than this score, was then assigned to corresponding individuals as the calibrated risk score. For instance, a patient got an initial risk score of 0.72 from the derived model while a total of 49 individuals attained their initial risk estimates ≥ 0.72 and 8 of them were cases, then the PPV (i.e., the calibrated risk score) aligned to 0.72 was calculated as 0.16 (8/49).

In the evaluation phase, the models built by different machine learning algorithms were applied to the validation cohort. Following that, we chose the one attaining the highest prediction accuracy (C-statistic) and computational efficiency as our final machine-learning-based predictive model. The ROC curve, sensitivities, as well as the PPVs were assessed. The stratified (low/intermediate/high) risk groups were assigned according to the relative risk, which was calculated by the calibrated risk score divided by population-based fall incidence rate. The intermediate-risk group was formed by a group of patients with averaged relative risk greater than 1 but less than 5, indicating a moderate risk of fall for the next year, while the high-risk category caught the individuals with averaged relative risk of fall equal to or greater than 5, indicating a much higher risk of fall for the next year comparing to the general population. The stratified fall-risk groups were further evaluated in the faller's population over their time of fall. After that, the odds ratios (ORs) and 95 % confidence intervals (CIs) were calculated for the captured predictive variable. The most

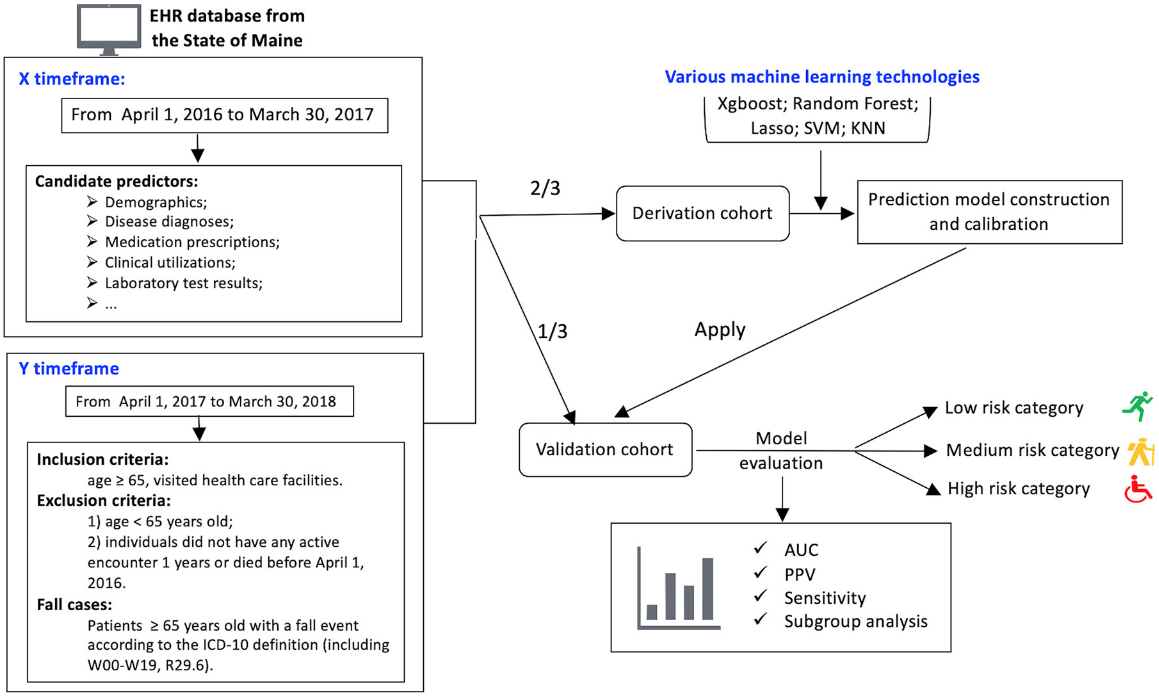
EHR database from the State of Maine
X timeframe:
From April 1, 2016 to March 30, 2017
Candidate predictors:
➢ Demographics;
➢ Disease diagnoses;
➢ Medication prescriptions;
➢ Clinical utilizations;
➢ Laboratory test results;
➢ ...
Y timeframe
From April 1, 2017 to March 30, 2018
Inclusion criteria:
age ≥ 65, visited health care facilities.
Exclusion criteria:
1) age < 65 years old;
2) individuals did not have any active encounter 1 years or died before April 1, 2016.
Fall cases:
Patients ≥ 65 years old with a fall event according to the ICD-10 definition (including W00-W19, R29.6).
2/3
1/3
Various machine learning technologies
Xgboost; Random Forest; Lasso; SVM; KNN
Derivation cohort
Prediction model construction and calibration
Apply
Validation cohort
Model evaluation
Low risk category
Medium risk category
High risk category
✓ AUC
✓ PPV
✓ Sensitivity
✓ Subgroup analysis

Fig. 1. Study design. The study cohort was derived from EHRs of patients with age of 65 years and older that visited Maine health care facilities and divided into the derivation and validation cohorts for model development and evaluation.

impactful predictors were then carefully investigated between the derived fall-risk categories (low, intermediate and high risk of fall in the following one year), revealing the clinical patterns and characteristics of patients in the high-risk category (i.e., disease comorbidity and concurrent medication therapies). In addition, the importance rank of the predictors was also assessed by the Lasso regression, uncovering the contribution of each predictor to the derived risk scores.

3. Results

3.1. Baseline characteristics

The dataset of older patients (≥65 years old) was stratified and divided into the derivation and validation cohorts in the ratio of 2:1, which comprised 176,816 and 88,409 patients, respectively. The baseline characteristics of the derivation and validation cohorts were listed in Table A. 1. Since the two cohorts were separated in a ratio, their baseline characteristics were all similarly distributed with no statistical difference, with the observed fall rates of 1.64 % for both cohorts. As a record of fall in EHR-based data essentially requires injury and subsequent visit to the doctor/hospital, those falls without injury would be missed in our study, leading to the much lower observed fall rate in our study than that from a general older population (i.e., 28%–35%).

3.2. Model performance

The fall risk prediction model was built on the derivation cohort and evaluated on the validation cohort. In summary, the XGBoost-based model had the best performance among various machine learning algorithms, attaining fitted and predicted C-statistics of 0.862 and 0.807 respectively (Fig. 2, Fig. A. 1). In the validation stage, with a total of 88,409 individuals, the developed fall-risk model classified 89.14 % (78,808/88,409) of them into the low-risk category, with a PPV as low as 1.04 %, whereas it identified 10.26 % (9075/88,409) and 0.59 % (526/88,409) of individuals into the intermediate and high risk groups of fall in the future one year, with PPVs of 6.31 % and 11.03 %, respectively (Table A. 2). Here, the PPV of a certain risk group was calculated as the number of cases divided by the number of members belong to this risk group, indicating the probability of falls in the risk group, while the relative risk was calculated as the PPV divided by population-based fall incidence rate, implying the relative risk of fall in this group compared to the general population. In the validation stage, the high and intermediate-risk category attained the relative risk of 6.7 and 3.9 while the low-risk category attained a much lower relative risk of less than 1.

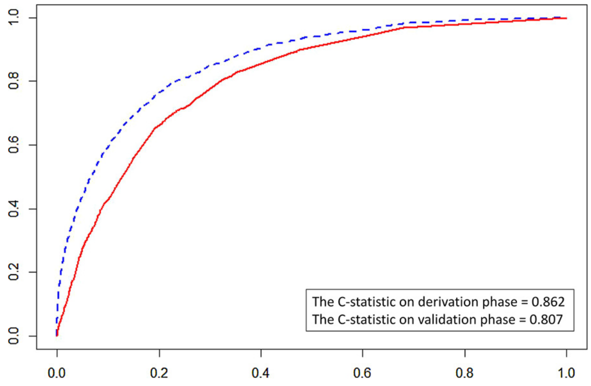

Fig. 2. The ROC curves derived from the derivation and validation stages based on the XGBoost algorithm.

In total, our approach successfully identified in advance 43.4 % of falls that happened all through the entire next year. When examining those short-term falls, our model captured more than 50 % of the events that happened within the first 80 days of the next year by classifying them into high and intermediate-risk categories (Fig. 3a). In particular, this percentage increased to 58.01 % and 54.93 % for the falls that happened within the first 30 and 30–60 days of next year. Furthermore, 50 % of high-risk and intermediate-risk true positives were confirmed to fall during the first 94 and 126 days of the next year respectively (Fig. 3b). On the contrary, those false-negative falls being classified into the low-risk category, tend to occur much later, with 65 % of which

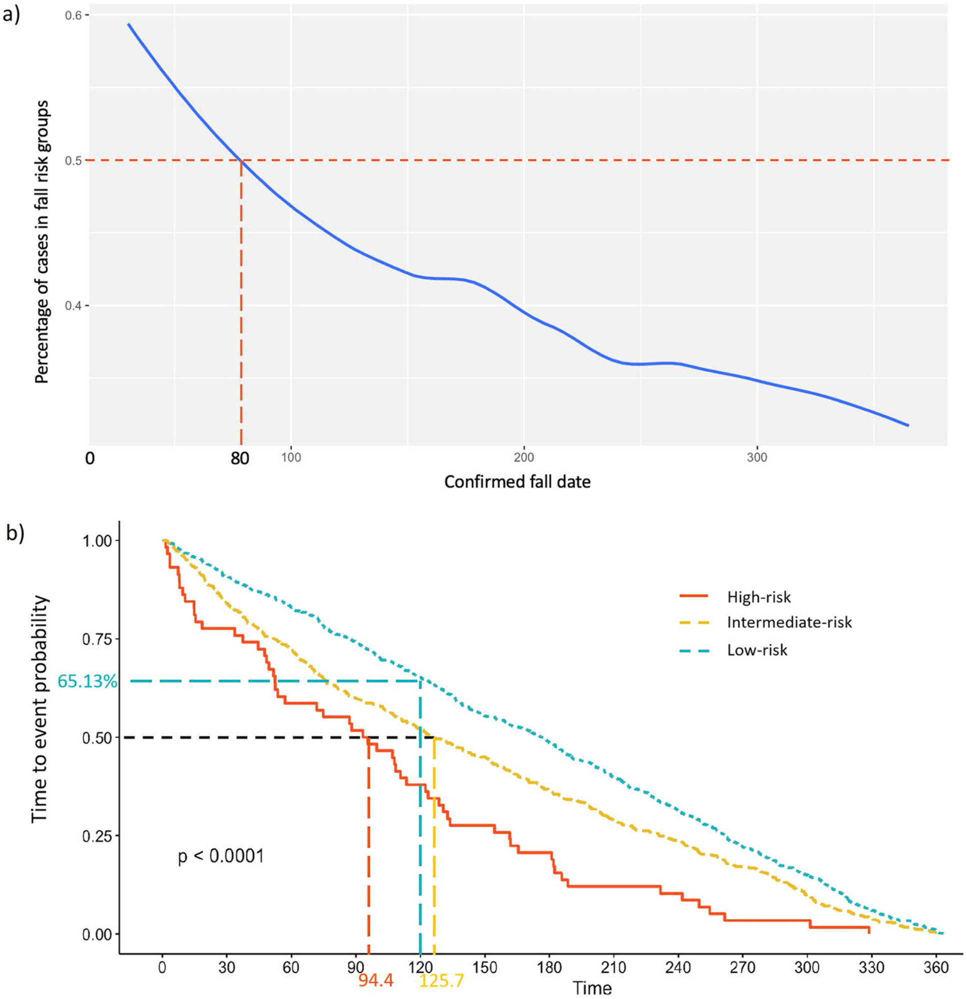

Fig. 3. a) Percentages of captured true-positive alters in the faller population, coordinated by the spectrum of their confirmed date of falls in the following one year (loess curve). b) The survival probability of fallers identified in three fall-risk categories over the following one year.

happened after the first 120 days. These findings indicated our model's promising prediction accuracy particularly for short-term falls.

Based on a pool of 10,198 predictor candidates, the XGBoost algorithm eventually captured 157 impactful features to form the final predictive model. These identified predictors were mainly demographic features (age and gender), chronic disease diagnoses, medication prescriptions and clinical utilization indicators. The most impactful 55 features are summarized in the Table A. 3, with their ORs or coefficients and 95 % CIs calculated from the univariate logistic regression in the validation cohort. Age was recognized as the strongest predictor of falls in the older people, attaining an OR of 6.07 in the validation cohort. Females were more likely to have a fall event than male in our dataset, which was consistent with previous findings [4].

3.3. Significant predictors

Almost half of predictors involved in the model were diagnoses of diseases (acute/chronic) and physical dysfunctions. The conditions with ORs > 3 were abnormalities of gait and mobility, Parkinson's disease (PD), cognitive diseases (neurodegenerative diseases, including Alzheimer's disease), orthostatic hypotension, cerebral infarction, heart failure and muscle disorders (Table A. 3). When compared to the low-risk category, the high-risk individuals were more likely to have the disease conditions of cognitive disease and abnormalities of gait and balance, with relative risks > 8, and resulting in a high rate of confirmed falls (> 12 %) in the future one year (Fig. 4a). Here the relative risk was calculated as the ratio of the probability of a certain disease occurring in the high-risk category versus the probability of that disease in the low-risk category. Another critical predictor, abrasion of knee, was not a quite common condition in the high-risk category (1%) but attained the highest relative risk (131.10) comparing to the low-risk category and also confirmed to have a high rate of fall in the following year (14 %). Diseases located on the right upper side of Fig. 4a all revealed their increased dominance in the high-risk category and the higher confirmed fall rate.

Another large group of predictors engaged in our model was medications for various diseases or conditions. Most were used to treat cardiovascular diseases, mental disease and epilepsy. Loop diuretic, beta-adrenergic blocker, antiepileptics attained ORs > 3, and selective serotonin reuptake inhibitors (SSRIs), atypical antipsychotic, cholinergic muscarinic antagonist, serotonin and norepinephrine reuptake inhibitor (SNRI) reached ORs > 2.5 in our validation cohort (Table A. 3). Patients treated by antiepileptics or tricyclic antidepressants (TCAs) were not only enriched in the high-risk category (i.e., relative risks of 5, calculated as the ratio of the probability of undertaking the medication in the high-risk category versus that probability in the low-risk category), but also attained the highest rate of confirmed fall in the next one year (16.41 % and 16.67 %, respectively), indicating their powerful predictive value (Fig. 4b). The high-risk category was also enriched by individuals with prescriptions of beta-adrenergic blocker and loop diuretic, opioid agonist and Nonsteroidal anti-inflammatory drugs (NSAID).

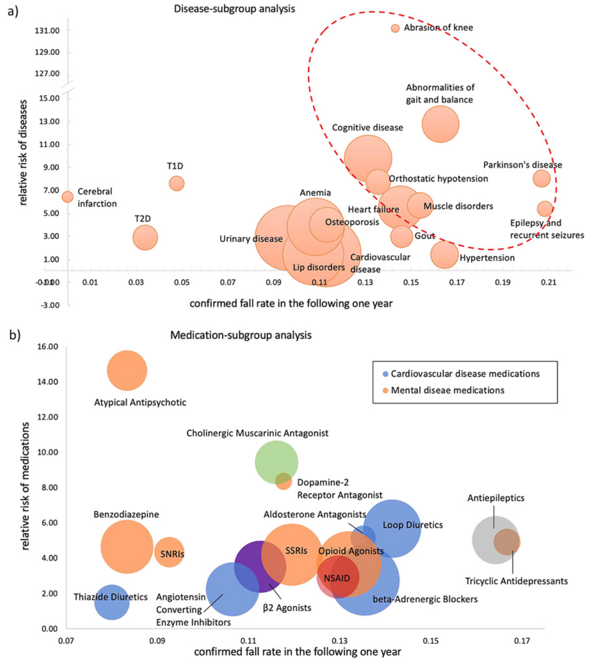

Fig. 4. Patients' relative risk of a) carrying certain diseases and b) undertaking certain medications in the high-risk category versus that in the low-risk category, coordinated by the confirmed fall rate of each sub-cohort in the high-risk category. These 17 diseases and 18 medications all significantly contributed to the built fall risk model. The percentage of patients in the high-risk category with each circumstance is denoted by the size of circles.

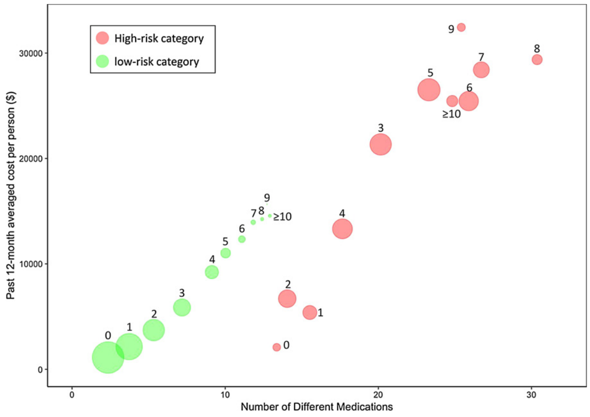

Fig. 5. The distribution of patients' average clinical costs in the past 12 months against their average number of distinct medication prescriptions for the defined disease comorbidity subgroups (determined by the counts of top 20 associated disease conditions a patient had). The size of the circles represents the percentage of patients with certain comorbidity in the validation cohort.

In our study, several EHR-derived disease comorbidity and clinical utilization indicators also revealed their predictive power in the fall risk model. They were counts of disease diagnoses, counts of medication consumptions, counts of outpatient/emergency admissions, lengths of inpatient stay, and total clinical cost during last year (Table A. 3). As Fig. 5 showed, over 50 % of the high-risk patients had severe disease comorbidity (≥5 fall-associated diseases) and also received an increased number of medications, leading to an inflation of their total healthcare costs, while only 7% of the low-risk patients had such severe level of disease comorbidity.

Fall history was also confirmed in our study as one of the most important predictors of future-one-year falls, reaching an OR of 5.27 in the validation cohort. In the Lasso regression that treated the derived fall-risk scores as outcome variable, fall history was recognized as the 4th important variable, and together with cognitive diseases, abnormalities of gait and balance, Parkinson's disease and osteoporosis, in the rank of priority, formed the top-5 contributors to the variation of risk scores (Table A. 4).

4. Discussion

4.1. *Summary of the study*

In this study, we constructed an EHR-based fall risk predictive model that adopted a machine-learning-based algorithm, XGBoost, to automatically integrate useful clinical information of disease diagnoses, medication consumption, clinical utilization, lab-test results and predicted an older individual's risk of fall in the future one year. In the validation phase, this model attained a C-statistic of 0.807, and stratified individuals into three distinct risk categories of fall (high, intermediate and low). About 43.4 % of the individuals that had a confirmed fall event in the future one year were classified into the increased risk categories. More importantly, our model successfully captured 58.01 % and 54.93 % of falls that happened within the first 30 and 30–60 days of next year into the risky group, respectively. 50 % of the identified high-risk true positives were confirmed to fall during the first 94 days of the next year, indicating the model's better performance for the short-term fall prediction.

4.2. *The machine learning algorithms*

The high-dimensional EHR data usually requires the algorithms to have the capacity of handling thousands of correlated clinical parameters simultaneously, where the number of parameters is usually much greater than the number of samples. Therefore, the traditional statistical methods may not be applicable due to their limitations such as reliance on assumptions, low computational efficiency and disability of handling high-dimensional data. Under this circumstance, the data-driven machine-learning approaches can be a good choice [21–23]. In this study, we included several popular linear and non-linear machine learning algorithms: RF is known to be robust to overfitting and correlated variables [24]; XGBoost is consistently rated as one of the best-performing machine learning algorithms nowadays [24]; Lasso is a linear algorithm and well-suited for sparse data setting when only a small number of variables would be valuable predictors in the model [25]; SVM scales relatively well in high dimensional data when the structure of the data is unknown; KNN was involved as a reference for its simplicity to implement and no requirement of data training process [26].

The results showed the XGBoost-based prediction model attained the highest prediction accuracy. The XGBoost algorithm has advantages of considering multiple potentially correlated predictors simultaneously, being able to handle possible underlying non-linear correlations (e.g., high-dimension interactions/correlations). In addition, the algorithm has high computational efficiency, and can provide variable importance table for model interpretation. As a member of tree-based modeling algorithm, XGBoost has been proven to have an innate ability to be robust to highly correlated variables [27]. We introduced a correlation screening process to remove redundant variables with spearman correlation > 0.7 before the modeling phase. With the survived 6949 variables, the XGBoost algorithm attained fitted and predicted C-statistics of 0.851 and 0.803, respectively. This finding with similar predictive performance to that of our original model without the screening (0.862 and 0.807), revealed that, elimination of the redundant variables from our EHR feeds would not have major impacts to improve XGBoost's predictive accuracy. Uniquely, XGBoost's tolerance of feature redundancy can train to give higher weights to a combination of highly predictive but also correlated features, and simultaneously provide additional clues for posterior risk interpretation. For instance, the two variables, previous-year distinct diagnoses and patient's medical cost, both were retained in our original prediction model with high weights. However, both variables are with a strong spearman correlation of 0.78, implying that the severe disease comorbidity and the subsequent increase of clinical utilization are instrumental for the prediction of the risk of fall.

For other used algorithms, RF is similar to XGBoost, but resulted in disappointing performance in our study using the default parameter settings (Fig. A. 1). Comparing to XGBoost, RF has less computational efficiency, preventing additional tuning of the parameters to improve the performance. As a linear prediction method, Lasso's low performance on the prediction may indicate non-linear associations between EHR-based fall risk variables. SVM's performance relies on the choice of kernel function, which may be chosen inappropriately in our case leading to the bias [28]. KNN's low accuracy may arise from its sensitivity to the large amount of noises in our high-dimensional EHR data [29].

In our study, the fall risk prediction model was constructed on the derivation cohort (2/3) and evaluated on the validation cohort (1/3). In addition, the K-fold cross validation, was also tried to evaluate the performance of XGBoost. Rather than using a pre-divided derivation set (2/3) for training and a validation set (1/3) for testing, our 10-fold cross-validation estimated the algorithm's performance by going through the entire dataset for both training and validation [28]. As a result, the mean of validated C-statistics in the 10-fold cross-validation attained a value of 0.815, slightly higher than our validated C-statistic (0.807) using the 1/3 testing cohort. Our K-fold cross validation results are in line with our training/testing results, supporting the robustness of our XGBoost model on prediction.

4.3. Implications of the findings

In our study, disorders related to cognitive impairments were the dominant group of diagnostic predictors, including Alzheimer's disease, amnesia, symptoms and signs of cognitive functions and awareness, and degenerative diseases of nervous system. Cognitive disorders could influence elders' attention, executive function, information processing and reaction time, cause functional dependency and disabilities, and lead to gait and balance problems [30]. Plenty of studies had revealed that the increased fall risk in PD patients was mainly affected by the PD-induced declined cognition, losing control/sensation of limbs, increased disability in many gait-dependent activities [31]. Other identified predictors, such as epilepsy, recurrent seizures, cardiovascular and cerebrovascular diseases, and lower extremity strength (e.g., abrasion of knee and muscle disorders) can induce gait and balance complications, which were recognized as triggers of fall in multiple studies [32,33]. Therefore, those identified disease conditions were strongly correlated with each other, and performed interactively to increase fall risk.

In terms of medications, psychotropic medications treating depression, bipolar and anxiolytics were proven to steadily increase the risk of fall in older adults [34], many of which were predictors in our built model, such as SNRIs, SSRIs, TCAs, benzodiazepines and so on. The captured cardiovascular medications contained diuretics, Angiotensin Converting Enzyme Inhibitors (ACE inhibitors), β-adrenergic blockers, β2 agonists and aldosterone antagonists, most of which were antihypertensive drugs. Our study and several population-based studies revealed a positive correlation between the antihypertensive medications and an increased risk of fall injuries [35,36], whereas others illustrated that ACE inhibitors or calcium channel blockers could reduce such risk [36]. Such controversial findings should be addressed by more advanced studies. In addition, the captured NSAID [37], cholinergic muscarinic antagonists [38] and antiepileptics [39] were recognized as fall risk factors for a long time, while the identified β2 agonists, as a medication to treat asthma and other pulmonary disorders, could cause headache, tremor and muscle cramps [40], and was also proven to be a risk factor of fall.

4.4. Utilization and benefits of the fall risk predictive model

In our study, the identified high-risk patients of fall have shown severe disease comorbidities, increased number of distinct medication prescriptions and much higher historical clinical costs. With the advantages of diverse and readily accessible data source, the EHR-based fall prediction could be an ideal and beneficial tool as the first step to a feasible fall prevention strategy. When those older adults with high risk of fall were identified, their personal unique risk factors would be captured, and their fall preventing strategies could be designed and proposed accordingly. For instance, educations to improve the awareness of fall risk, minimization or withdrawal of specific psychoactive or cardiovascular medications, detailed exercise therapies for patients with balance and gait issues. It is hoped that, by implementing the fall prediction procedure and the corresponding interventions, the fall incidence rate could be reduced and the quality of life in older population could be improved eventually.

As a limitation, our EHR-based dataset suffered from under-reporting issue for falls without injury, resulting in a much lower fall rate than that in the community-dwelling older population. Furthermore, the mental illness diagnoses have been masked in our dataset for privacy protection in Maine. As a result, the confounding effect between the masked psychotic disorders and the used medication cannot be directly ascertained in our study. Another limitation is that, the medication consumptions were binarily coded in our study, while the dosage information was not considered, which should be taken into account in the future studies.

5. Conclusion

In conclusion, we have constructed and validated a powerful risk assessment tool to predict older adults' risk of fall in the future one year, by using the EHR data from the older population in Maine. We hope that this constructed fall risk assessment tool could be immediately deployed to provide early warnings to older adults with increased fall risk and identifying their personalized risk factors to facilitate customized fall interventions.

Summary points

What is already known:

Traditional fall risk assessment tools are mainly physical function evaluations that monitor an individual's static and dynamic gait and balance performance.

Electronic health records have been recognized as a good source for disease management and disease risk prediction.

Most existing EHR-based fall risk models were developed using traditional statistical approaches, such as logistic regressions or cox regressions.

What this study has added:

By using the advanced machine-learning algorithms, this new risk assessment tool attained an improved discriminative ability from the statewide electronic health records and can alarm automatically for elders at increased risk of falls during the next one year.

This new tool successfully captured high-risk fallers with conditions of severe disease comorbidities, an enrichment of fall-increasing cardiovascular or mental drugs and increased historical clinical utilization.

This risk assessment tool can be immediately deployed in the electronic health system to provide early warnings, recognize personalized risk factors and facilitate customized fall interventions.

Funding

CY is supported by the National Natural Science Foundation of China (Award No.81402762).

Author's statement

The authors certify that this manuscript is not under review by any other journal. All authors have contributed to and read the manuscript and approved the final copy.

Declaration of Competing Interest

The authors certify that they do not have any financial or other conflicts of interest in relation to this manuscript.

Acknowledgments

The authors would like to thank and express their gratitude to the hospitals, medical practices, physicians, and nurses participating in Maine's HIE. They also thank the biostatistics colleagues at the Department of Health Research and Policy for critical discussions.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104105.

References

[1] S. Yoshida, A global report on falls prevention epidemiology of falls, WHO Rep. 2012 (2012) 5–8.

[2] F. Baranzini, M. Diurni, F. Ceccon, N. Poloni, S. Cazzamalli, C. Costantini, et al., Fall-related injuries in a nursing home setting: is polypharmacy a risk factor? BMC Health Serv. Res. 9 (December) (2009) 228.

[3] C.S. Florence, G. Bergen, A. Atherly, E. Burns, J. Stevens, C. Drake, Medical costs of fatal and nonfatal falls in older adults, J. Am. Geriatr. Soc. 66 (April (4)) (2018) 693–698.

[4] A.F. Ambrose, G. Paul, J.M. Hausdorff, Risk factors for falls among older adults: a review of the literature, Maturitas 75 (1) (2013) 51–61.

[5] L. Letts, J. Moreland, J. Richardson, L. Coman, M. Edwards, K.M. Ginis, et al., The physical environment as a fall risk factor in older adults: systematic review and meta-analysis of cross-sectional and cohort studies, Aust. Occup. Ther. J. 57 (February (1)) (2010) 51–64.

[6] J. Howcroft, E.D. Lemaire, J. Kofman, W.E. McIlroy, Elderly fall risk prediction using static posturography, PLoS One 12 (2) (2017) e0172398.

[7] A. Borowicz, E. Zasadzka, A. Gaczkowska, O. Gawlowska, M. Pawlaczyk, Assessing gait and balance impairment in elderly residents of nursing homes, J. Phys. Ther. Sci. 28 (September (9)) (2016) 2486–2490.

[8] L. Hou, H. Yu, Y. Ma, L. Li, X. Chen, L. Wang, et al., A screening tool using five risk factors was developed for fall-risk prediction in Chinese community-dwelling elderly individuals, Rejuvenation Res. 21 (5) (2017) 416–422.

[9] P. Palumbo, J. Klenk, L. Cattelani, S. Bandinelli, L. Ferrucci, K. Rapp, et al., Predictive performance of a fall risk assessment tool for community-dwelling older people (FRAT-up) in 4 European cohorts, J. Am. Med. Dir. Assoc. 17 (December (12)) (2016) 1106–1113.

[10] J.Y. Lee, Y. Jin, J. Piao, S.M. Lee, Development and evaluation of an automated fall risk assessment system, Int. J. Qual. Health Care 28 (2) (2016) 175–182.

[11] A. Marier, L.E.W. Olsho, W. Rhodes, W.D. Spector, Improving prediction of fall risk among nursing home residents using electronic medical records, J. Am. Med. Inform. Assoc. 23 (2) (2016) 276–282.

[12] S. Yokota, K. Ohe, Construction and evaluation of FiND, a fall risk prediction model of inpatients from nursing data, J. Nurs. Sci. 13 (April (2)) (2016) 247–255.

[13] A. Baus, J. Coben, K. Zullig, C. Pollard, C. Mullett, H. Taylor, et al., An electronic health record data-driven model for identifying older adults at risk of unintentional falls, Perspect. Heal Inf. Manage. 14 (Fall) (2017) 1b.

[14] L. Kang, X. Chen, P. Han, Y. Ma, L. Jia, L. Fu, et al., A screening tool using five risk factors was developed for fall-risk prediction in chinese community-dwelling elderly individuals, Rejuvenation Res. 21 (October (5)) (2018) 416–422.

[15] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[16] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet], New York, NY, USA: ACM, 2016, pp. 785–794, , https://doi.org/10.1145/2939672.2939785 (KDD' 16). Available from:.

[17] J. Friedman, T. Hastie, N. Simon, R. Glmnet, Lasso and Elastic-Net Regularized Generalized Linear Models, R package version 2.0 (2016).

[18] B.D. Ripley, k-Nearest Neighbour Classification, R package version 7.4 (2015).

[19] A. Karatzoglou, A. Smola, K. Kurt, K. Hornik, kernlab: Kernel-Based Machine Learning Lab, R package version 0.9 (2018).

[20] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning (2005).

[21] A.J. Steele, S.C. Denaxas, A.D. Shah, H. Hemingway, N.M. Luscombe, Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, PLoS One 13 (8) (2018) e0202344.

[22] C. Ye, O. Wang, M. Liu, L. Zheng, M. Xia, S. Hao, et al., A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data, J. Med. Internet Res. 21 (7) (2019) e13719.

[23] C. Ye, T. Fu, S. Hao, Y. Zhang, O. Wang, B. Jin, et al., Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning, J. Med. Internet Res. 20 (1) (2018) e22.

[24] M. Hodnett, J.F. Wiley, R Deep Learning Essentials: A Step-by-step Guide to Building Deep Learning Models Using TensorFlow, Keras, and MXNet, 2nd ed., Packt Publishing, 2018 380 p..

[25] C.S. Signorino, A. Kirchner, Using LASSO to model interactions and nonlinearities in survey data, Surv. Pract. 11 (1) (2018).

[26] S. Neelamegam, E. Ramaraj, Classification algorithm in data mining : an overview, Int. J. P2P Netw. Trends Technol. 3 (5) (2013).

[27] Tianqi Chen, T.H. Michaël Benesty, Understand Your Dataset With Xgboost. [Internet]. R Project, Available from: (2018) https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html#numeric-v.s.-categorical-variables.

[28] J. Mahjoobi, E. Adeli Mosabbeb, Prediction of significant wave height using regressive support vector machines, Ocean Eng. 36 (5) (2009) 339–347.

[29] P. Cunningham, S.J. Delany, K -Nearest neighbour classifiers, Mult. Classif. Syst. 4 (2007) 1–17.

[30] N.B. Alexander, J.M. Hausdorff, Guest editorial: linking thinking, walking, and falling, J. Gerontol. 63 (2008) 1325–1328.

[31] C.G. Canning, S.S. Paul, A. Nieuwboer, Prevention of falls in Parkinson's disease: a review of fall risk factors and the role of physical interventions, Neurodegener. Dis. Manage. 4 (3) (2014) 203–221.

[32] S. Deandrea, E. Lucenteforte, F. Bravi, R. Foschi, C. La Vecchia, E. Negri, Risk factors for falls in community-dwelling older people: a systematic review and meta-analysis, Epidemiology 21 (September (5)) (2010) 658–668.

[33] M.E. Tinetti, M. Speechley, S.F. Ginter, Risk factors for falls among elderly persons living in the community, N. Engl. J. Med. 319 (December (26)) (1988) 1701–1707.

[34] S. Hartikainen, E. Lonnroos, K. Louhivuori, Medication as a risk factor for falls: critical systematic review, J. Gerontol. A Biol. Sci. Med. Sci. 62 (October (10)) (2007) 1172–1181.

[35] J. Gribbin, R. Hubbard, J.R.F. Gladman, C. Smith, S. Lewis, Risk of falls associated with antihypertensive medication: population-based case-control study, Age Ageing 39 (September (5)) (2010) 592–597.

[36] G. Zang, Antihypertensive drugs and the risk of fall injuries: a systematic review and meta-analysis, J. Int. Med. Res. 41 (October (5)) (2013) 1408–1417.

[37] L.R. Findley, M.N. Bulloch, Relationship between nonsteroidal anti-inflammatory drugs and fall risk in older adults, Consult. Pharm. 30 (June (6)) (2015) 346–351.

[38] K. Ruxton, R.J. Woodman, A.A. Mangoni, Drugs with anticholinergic effects and cognitive impairment, falls and all-cause mortality in older adults: a systematic review and meta-analysis, Br. J. Clin. Pharmacol. 80 (August (2)) (2015) 209–220.

[39] Y. Haasum, K. Johnell, Use of antiepileptic drugs and risk of falls in old age: a systematic review, Epilepsy Res. 138 (December) (2017) 98–104.

[40] F. de Vries, S. Pouwels, M. Bracke, H.G.M. Leufkens, C. Cooper, J.-W.J. Lammers, et al., Use of beta-2 agonists and risk of hip/femur fracture: a population-based case-control study, Pharmacoepidemiol. Drug Saf. 16 (June (6)) (2007) 612–619.