Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits

Brian W. Patterson, MD, MPH,*† Collin J. Engstrom, MS,‡ Varun Sah, MS,‡ Maureen A. Smith, MD, PhD,‡∥ Eneida A. Mendonça, MD, PhD,¶ Michael S. Pulia, MD, MS,* Michael D. Repplinger, MD, PhD,* Azita G. Hamedani, MD, MBA, MPH,* David Page, PhD,‡∥ and Manish N. Shah, MD, MPH*§**

Background: Machine learning is increasingly used for risk stratification in health care. Achieving accurate predictive models do not improve outcomes if they cannot be translated into efficacious intervention. Here we examine the potential utility of automated risk stratification and referral intervention to screen older adults for fall risk after emergency department (ED) visits.

Objective: This study evaluated several machine learning methodologies for the creation of a risk stratification algorithm using electronic health record data and estimated the effects of a resultant intervention based on algorithm performance in test data.

Methods: Data available at the time of ED discharge were retrospectively collected and separated into training and test datasets. Algorithms were developed to predict the outcome of a return visit for fall within 6 months of an ED index visit. Models included random forests, AdaBoost, and regression-based methods. We evaluated models both by the area under the receiver operating characteristic (ROC) curve, also referred to as area under the curve (AUC), and by projected clinical impact, estimating number needed to treat (NNT) and referrals per week for a fall risk intervention.

Results: The random forest model achieved an AUC of 0.78, with slightly lower performance in regression-based models. Algorithms with similar performance, when evaluated by AUC, differed when placed into a clinical context with the defined task of estimated NNT in a real-world scenario.

Conclusion: The ability to translate the results of our analysis to the potential tradeoff between referral numbers and NNT offers decisionmakers the ability to envision the effects of a proposed intervention before implementation.

Key Words: falls, screening, electronic health record, machine learning, emergency medicine

ORIGINAL ARTICLE

From the *BerbeeWalsh Department of Emergency Medicine, University of Wisconsin School of Medicine and Public Health; †Health Innovation Program; ‡Department of Computer Sciences, University of Wisconsin-Madison; Departments of §Population Health Sciences; ¶Family Medicine; ¶Biostatistics and Medical Informatics; #Pediatrics, University of Wisconsin School of Medicine and Public Health; and **Department of Medicine, Division of Geriatrics and Gerontology, University of Wisconsin School of Medicine and Public Health, Madison, WI.

Supported by funding from the Agency for Healthcare Research and Quality (AHRQ), grant numbers: K08HS024558 (B.W.P.) and K08HS024342 (M.S.P.), as well as the National Institutes of Health (NIH), grant numbers K08DK111234 (M.D.R.) and K24AG054560 (M.N.S.). The research was also supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the NIH.

The authors declare no conflict of interest.

Reprints: Brian W. Patterson, MD, MPH, University of Wisconsin School of Medicine and Public Health, 800 University Bay Drive, Suite 310, Mail Code 9123, Madison, WI 53705. E-mail: bpatter@medicine.wisc.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s website, www.lww-medicalcare.com.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 0025-7079/19/5707-0560

Medical Care • Volume 57, Number 7, July 2019

560 | www.lww-medicalcare.com

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.
risk stratification for a particular scenario, such as ruling out a rare disease, confirming a particular diagnosis, or reducing population risk via an intervention—in this case, referral for a fall risk reduction intervention.

Such an intervention already exists at our institution in the form of a multidisciplinary falls clinic. On the basis of prior literature, we estimate a relative risk reduction of 38% for future falls for patients enrolled in such a program. Currently, very few referrals are made to the falls clinic from the ED. Before initiating an automated referral program, decisionmakers must understand both the anticipated number of referrals generated and the effectiveness of such referrals in preventing future falls. To do so, decisionmakers may be better served by extrapolations of a model’s performance in a given population than by test characteristics such as AUC. This information would allow a clinical site to select the most appropriate risk stratification algorithm, and most appropriate threshold point, to maximize patient benefit within the constraints of available resources and acceptable effectiveness. In this study, we developed several machine learning models to predict 6-month fall risk after an ED visit. We evaluated these models both using AUC analysis and by interpreting model performance to describe potential clinical tradeoffs more concretely in terms of referrals per day and numbers needed to treat (NNT) for prevention of a fall.

METHODS

Study Design and Setting

We performed a retrospective observational study using patient EHR data at a single academic medical center ED with level 1 trauma center accreditation and ~60,000 patient visits per year. The goal of developing the models was to create an alert at the time of an ED visit suggesting referral of patients who are at heightened risk of fall for an existing multidisciplinary falls intervention. In our case, based on discussions with our falls clinic, an estimated 10 referrals per week were seen as operationally feasible. Using the available EHR data, we created risk-stratification models for fall revisits to the ED. Our outcome of interest was a fall visit to the same ED in the 6 months after an index visit. Although this paper focuses on predicting fall revisits, the methodology we describe is robust and lends itself to any clinical risk stratification prediction task.

Data Selection and Retrieval

EHR data for patients aged 65 years and older who visited the study ED were acquired for a duration of 43 months starting January 2013, with an additional 6 months of follow-up data collected for outcome determination. Available EHR features were evaluated for inclusion under the conceptual framework of the Andersen Behavioral Model of Health Services Use, a well-established model which provides a context for characterizing the many factors which lead to health care utilization. This model has been used to frame numerous prior studies involving ED use and falls among older adults. For each visit, discrete data available within the EHR at the time of the ED visit were collected to create data features including patient demographics, historical visits, and visit patterns and diagnoses, as well as visit-specific information including timing, laboratory tests performed and results thereof, vital signs, chief complaint, and discharge diagnoses. Features were selected based on their availability, clinical relevance, and potential to provide predictive value for fall revisit risk estimation. Another important criteria for feature selection were to exclude attributes that contained information obtained after an index visit.

The data were organized and analyzed at the level of an ED visit (as opposed to patient level) as our objective was to stratify risk for a fall revisit based on index visit data alone. Visits by patients who were transferred from other health care facilities were rejected as part of our primary exclusion criteria. We excluded visits that resulted in hospital admissions, as our algorithm would only be implemented for patients who were discharged from the ED. Finally, we excluded patients who did not have a primary care provider (PCP) in our network, as our intervention was specifically aimed towards referring in-network patients. At the end of the exclusion procedures, we were left with 10,030 records.

Feature Preparation

The encoding process for features depended on whether they were numerical or categorical in nature. Numerical features such as age, vital signs during the index ED visit, duration of the index visit, and a number of primary care or hospital visits in the 6 months before the index visit were treated as continuous values. Attributes related to Elixhauser comorbidity index, Hendrich II score, patients’ demographics, medications, and laboratory results were treated as categorical variables. In the case of numerical features, we dropped records that had missing values due to the relatively small number of records that were incomplete in this regard, which left us with 9687 records. However, for categorical variables, missing values were considered as a separate category—in general, the absence of most categorical features could be potentially informative for decision making by the predictive models. At the end of the feature engineering process, we obtained our final dataset which was comprised of 725 features. The feature preparation phase was completely independent of outcome status.

Model Development

Once our features were selected and prepared, we created predictive models from the data. We tested several regression-based methodologies, including thresholded linear regression and logistic regression, both unregularized and including lasso and ridge penalties. We also included 2 tree-based methodologies: random forests and AdaBoost. Appendix A (Supplemental Digital Content 1, http://links.lww.com/MLR/B801) provides a nontechnical description of the methods used. Models were generated using the Scikit-learn package in Python. The dataset created at the end of feature preparation was split into training and test sets in a 3:1 ratio. We split data chronologically, with the final 25% of visits kept as a holdout test set, and the earliest 75% of data retained as a training set. The training set was further split, again chronologically in a 3:1 ratio, to create a tuning set for interim validation.

Models were initially trained on the smaller training set, where tunable parameters were varied using a grid search pattern to achieve best results within the tuning set. Finally, we picked the 6 models that performed best on the tuning set and trained 1 model of each type on the entire training set. These models were then evaluated on the test data that had been held out during the
previous phase. As our dataset was skewed, with more patients who did not fall than those who did, we upsampled the positive class records while training models to provide a weighting effect to incentivize correct classification of fall cases. This was achieved by randomly duplicating positive cases in the training set until their frequency equaled that of negative cases. Upsampling was carried out only after the training set had been split into a tuning set, to ensure that no duplicate records created as a result of upsampling on the entire training set were members of both the training and tuning set. Further, the test set was not subjected to any upsampling, to maintain the true population distribution in the evaluation set to simulate performance assessment on future data.

Model Evaluation

Our initial evaluation of the trained models involved comparing the AUC. 95% confidence intervals (CIs) were generated in STATA (StataCorp., College Station, TX) using a nonparametric bootstrapping with the Rocreg command and 1000 iterations. We then generated classification statistics for each model at each potential threshold value, consisting of performance within the evaluation set in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). We were able to use this data to extrapolate both referrals per week and NNT. Estimated referrals per week were calculated by taking the total percentage of TP and FP results (all patients flagged “positive”) at a given threshold from each model and multiplying by the weekly visit volume. NNT was estimated by assuming that the falls reduction clinic would provide a relative risk reduction of 38% (95% CI: 21%–52%) based on the results of the PROFET randomized clinical trial which studied a similar intervention in practice and clinic would provide a relative risk reduction of 38% (95% CI: 21%–52%).

Relative risk reduction and CIs were generated from the reported PROFET data using STATA. The absolute fall risk for a population of patients above a given risk threshold in our models was calculated as the ratio of TP (patients who did not fall than those who did) to the total number of positive cases. This absolute risk was multiplied by the relative risk reduction of 0.38 to estimate an absolute risk reduction, and the inverse of the absolute risk reduction was taken to generate the NNT. For instance; if the absolute fall risk in the flagged positive group was 60%, the estimated NNT was $1/(0.38 \times 0.6) = 4.4$ referrals per fall prevented. These projected performance measures were used to create plots that visually described the tradeoff between risk reduction gained per referral and number of referrals expected per day. We have created a toolkit for recreating these plots from ROC data, which is available at https://www.hipxchange.org/NNT.

RESULTS

We had 32,531 visits to the ED during the study period by adults aged 65 and older, of which 9687 were both discharged and had a PCP in our network and full numerical data, making up our study population (Fig. 1). Within this population, 857 patients returned within 6 months for a fall-related visit; the overall return rate for fall within 6 months was 8.8%. Demographics of patients by the outcome are presented in Table 1. As compared with patients who did not return for falls, those with falls were similar with regards to sex and insurance status but were older, more likely to have fallen on their index visit, and more likely to have been brought to the ED by an ambulance.

When comparing models based on AUC, the random forest model achieved an AUC of 0.78 (95% CI: 0.74–0.81), and AdaBoost also had an AUC of 0.78 (95% CI: 0.74–0.81). These tree-based models were the highest performers, followed by ridge-penalized logistic regression at 0.77 (95% CI: 0.73–0.80), lasso-penalized logistic regression at 0.76 (95% CI: 0.73–0.80), unpenalized linear regression at 0.74 (95% CI: 0.71–0.78), and unpenalized logistic regression at 0.72 (95% CI: 0.68–0.76). Figure 2 shows AUC plots for all tested machine learning models. Appendix B (Supplemental Digital

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No (%)</th>
<th>Visits Without 180-Day Return For Fall</th>
<th>Visits With 180-Day Return For Fall</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>9687</td>
<td>8830</td>
<td>857</td>
</tr>
<tr>
<td>Age [mean (SD)]</td>
<td>76.0 (8.4)</td>
<td>75.7 (8.3)</td>
<td>79.3 (8.9)</td>
</tr>
<tr>
<td>Female</td>
<td>5863 (60.5)</td>
<td>5286 (59.9)</td>
<td>577 (67.3)</td>
</tr>
<tr>
<td>White race</td>
<td>8980 (92.7)</td>
<td>8187 (92.7)</td>
<td>793 (92.5)</td>
</tr>
<tr>
<td>Medicare</td>
<td>8444 (87.2)</td>
<td>7705 (87.3)</td>
<td>739 (86.2)</td>
</tr>
<tr>
<td>Commercial/compensation</td>
<td>1210 (12.5)</td>
<td>1095 (12.4)</td>
<td>115 (13.4)</td>
</tr>
<tr>
<td>Other/self-pay</td>
<td>26 (0.3)</td>
<td>23 (0.3)</td>
<td>3 (0.4)</td>
</tr>
<tr>
<td>Mode of arrival</td>
<td>6641 (68.6)</td>
<td>6263 (70.9)</td>
<td>378 (44.1)</td>
</tr>
<tr>
<td>Fall at index visit</td>
<td>1543 (15.9)</td>
<td>1267 (14.4)</td>
<td>272 (31.7)</td>
</tr>
</tbody>
</table>

FIGURE 1. Patient allocation. Once the study population was defined; it was split at a 3:1 ratio into training and test sets. The training set was further split to create an intermediate tuning set. The testing set was then subjected to both bootstrapping with the Rocreg command and 1000 iterations. We had 32,531 visits to the ED during the study period by adults aged 65 and older, of which 9687 were both discharged and had a PCP in our network and full numerical data, making up our study population (Fig. 1). Within this population, 857 patients returned within 6 months for a fall-related visit; the overall return rate for fall within 6 months was 8.8%. Demographics of patients by the outcome are presented in Table 1. As compared with patients who did not return for falls, those with falls were similar with regards to sex and insurance status but were older, more likely to have fallen on their index visit, and more likely to have been brought to the ED by an ambulance.

When comparing models based on AUC, the random forest model achieved an AUC of 0.78 (95% CI: 0.74–0.81), and AdaBoost also had an AUC of 0.78 (95% CI: 0.74–0.81). These tree-based models were the highest performers, followed by ridge-penalized logistic regression at 0.77 (95% CI: 0.73–0.80), lasso-penalized logistic regression at 0.76 (95% CI: 0.73–0.80), unpenalized linear regression at 0.74 (95% CI: 0.71–0.78), and unpenalized logistic regression at 0.72 (95% CI: 0.68–0.76). Figure 2 shows AUC plots for all tested machine learning models. Appendix B (Supplemental Digital

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No (%)</th>
<th>Visits Without 180-Day Return For Fall</th>
<th>Visits With 180-Day Return For Fall</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>9687</td>
<td>8830</td>
<td>857</td>
</tr>
<tr>
<td>Age [mean (SD)]</td>
<td>76.0 (8.4)</td>
<td>75.7 (8.3)</td>
<td>79.3 (8.9)</td>
</tr>
<tr>
<td>Female</td>
<td>5863 (60.5)</td>
<td>5286 (59.9)</td>
<td>577 (67.3)</td>
</tr>
<tr>
<td>White race</td>
<td>8980 (92.7)</td>
<td>8187 (92.7)</td>
<td>793 (92.5)</td>
</tr>
<tr>
<td>Medicare</td>
<td>8444 (87.2)</td>
<td>7705 (87.3)</td>
<td>739 (86.2)</td>
</tr>
<tr>
<td>Commercial/compensation</td>
<td>1210 (12.5)</td>
<td>1095 (12.4)</td>
<td>115 (13.4)</td>
</tr>
<tr>
<td>Other/self-pay</td>
<td>26 (0.3)</td>
<td>23 (0.3)</td>
<td>3 (0.4)</td>
</tr>
<tr>
<td>Mode of arrival</td>
<td>6641 (68.6)</td>
<td>6263 (70.9)</td>
<td>378 (44.1)</td>
</tr>
<tr>
<td>Fall at index visit</td>
<td>1543 (15.9)</td>
<td>1267 (14.4)</td>
<td>272 (31.7)</td>
</tr>
</tbody>
</table>
The various machine learning models tested in this study differed in their ability to predict falls, with the random forest and AdaBoost models offering the best overall performance with an AUC of 0.78. On the basis of AUC alone, penalized regression-based models including ridge-penalized logistic regression offered similar performance with an AUC of 0.77. This result is consistent with other studies evaluating the performance of tree-based algorithms alongside regression-based methodologies.\textsuperscript{16,18,37–39} As opposed to traditional methods, tree-based methodologies have an improved ability to deal with complex variable interactions and nonlinear effects in large databases, which may explain their advantage in these instances.\textsuperscript{40}

When translating the models into potential deliverable performance at individual thresholds, the random forest-based approach offers the best performance in terms of NNT versus referrals in the proposed operational scenario, offering the ability to refer 10 patients per week at an NNT of 12.4 referrals to reduce the risk of an ED revisit for fall. Although these data are technically inferable based on the shape of the receiver operating characteristic curves, the degree of distinction between the models would likely not be apparent based on visual inspection alone to a reader not already expert in machine learning or statistics.

Algorithms derived by machine learning have become increasingly common in medicine, with significant excitement surrounding their potential to improve the ability to risk stratify patients.\textsuperscript{31,32} Unfortunately, gaps still exist between the ability to predict a potentially avoidable event and specific actionable interventions.\textsuperscript{33} In the majority of studies evaluating machine learning techniques, model performance is reported based on AUC or test characteristics such as sensitivity and specificity.\textsuperscript{44} These test characteristics may be useful for establishing predictive performance generally but may be misleading when not set into clinical context.\textsuperscript{21} Once AUC curves have been generated for a given risk stratification model in test data, calculating additional information including NNT and anticipated referrals requires only an algebraic transformation of the data, as long as a proposed intervention has been identified along with estimated effectiveness. The curves generated for this study communicate this tradeoff to policymakers, and provide a basis for comparison of anticipated “real-world” effects of model performance.

For any particular harm reduction intervention, there is a tradeoff when choosing a risk cutoff for a referral. The most total harm-reduction would be accomplished by simply referring all patients in a given population, however, such nonspecific referral would be costly in terms of time and resource use, and inefficient as many low-risk patients would receive minimal benefit, or potentially be exposed to risks of intervention. At the same time, selecting only those patients who are at extremely high risk of harm reduces the overall potential benefit of a risk reduction strategy by not offering it to a large proportion of patients who will go on to have the outcome of interest. In our example, where a set number of referral slots per week were available, and the task was to select the highest risk patients to fill those slots, the random forest algorithm was the best performer. If there had been only 5 referral spots available, however, the ridge-penalized logistic regression model would have been the top performer. Despite an overall slightly lower AUC, the ridge-penalized model had better performance in selecting those 5 patients at highest risk, achieving an NNT of \(\sim\)10 versus \(\sim\)12 for the tree-based models. If the intervention tied to the algorithm were a referral to a less resource-intensive community-based falls prevention program with more availability, policymakers may be looking in a region of higher referrals per week and higher NNT—in this region, model performance was generally similar between the various models.

The projections of performance generated in this study were based on model performance on a set of test data which immediately followed the training data chronologically. Although these projections are expected to help policymakers envision potential operational performance, they are not intended to replace the evaluation of performance during and after implementation. Machine learning models are tuned to specific population...
parameters, and subject to calibration drift as patient and data characteristics change over time, necessitating continued post-implementation monitoring to ensure effective results.

To our knowledge, 3 ED-specific fall screening instruments have been examined: Carpenter et al. examined a number of factors for association with future falls, proposing a screen of 4 independent factors, reporting a 4% probability of falling in their lowest risk group and 42% among the highest. Tiedemann et al. developed and externally validated a screening instrument with an AUC of 0.70, and Greenberg et al. utilized modified CAGE criteria but did not report fall outcomes in their pilot. As compared with these prior efforts, the machine learning-derived algorithms here offer improved performance in terms of test characteristics, and the advantage

FIGURE 3. Number needed to treat (NNT) versus anticipated referrals per week. This line shows the tradeoff between rising NNT and a rising number of potential referrals as a lower threshold for risk is selected within the model. The square represents a potential scenario in which all patients are referred regardless of model risk. The triangle represents the performance of a model with perfect discrimination (one which only refers patients who would definitely fall in the future and no one else). Error bars represent the 95% confidence interval of the relative risk estimation.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.
of not requiring the devotion of scant ED resources to in-person screening.42

Limitations

When generating our NNT we assumed that the relative risk reduction generated by our proposed intervention would remain constant across varying absolute risks. This assumption, while broadly made in medical decision making literature, is a simplification that is often, but not always, true.47,48 Furthermore, for the sake of simplifying our calculations, we assumed that all patients referred for fall intervention would attend the required intervention. If an estimate of the likelihood of completed referral were available, it could be taken into account in the NNT calculation.

We presented our NNT versus anticipated referrals per week curves with error bars based on the effectiveness estimate from the PROFET trial. PROFET measured the effectiveness of an intervention similar to our own falls clinic, but on a somewhat different outcome (any reported fall vs. ED visit for fall) and with somewhat different inclusion criteria (only selected older adults reporting to the ED for fall as opposed to all older adults). Given the relatively wide CI of the PROFET results, we feel the included error bars provide a reasonable estimate of uncertainty, however, these could be widened to incorporate the estimated impact of other sources of potential variation in predicted effectiveness.

During model development, we chose to censor visits which were missing data features encoded as continuous variables (categorical variables were encoded to allow a “missing” category). Although the inclusion of only complete records has the potential to introduce bias,49 only 343 (3%) of records were dropped for incompleteness, suggesting the minimal potential for change in algorithm performance if this data were imputed.

Our model was trained on an outcome of return visits to our ED for falls. Patients who fell may in some instances have presented to other EDs, in which case they were not captured to our ED for falls. Patients who fell may in some instances have presented to other EDs, in which case they were not captured.

Our model was trained on an outcome of return visits to our ED for falls. Patients who fell may in some instances have presented to other EDs, in which case they were not captured.

CONCLUSIONS

In this analysis, we developed an algorithm which had an AUC of 0.78 for prediction of a return visit to the ED for fall within 6 months of an index visit. Placed in the clinical context of harm reduction, this offered the ability to refer 10 patients per week to our fall clinic with a predicted NNT of 12 referrals to reduce the risk of a single fall. Our ability to translate the results of our analysis to the potential tradeoff between referral numbers and NNT offers decisionmakers the ability to envision the effects of a proposed intervention before implementation.

REFERENCES

Psychometric Evaluation of an Instrument to Measure Prospective Pregnancy Preferences: The Desire to Avoid Pregnancy Scale: Erratum

In the February 2019 issue of Medical Care, there was an error in the article “Psychometric Evaluation of an Instrument to Measure Prospective Pregnancy Preferences: The Desire to Avoid Pregnancy Scale.” On page 156, the sentence “As hypothesized, women not using contraception scored significantly higher on the DAP than those using a method (0.75 logits higher) and those who had not had sex in 30 days (0.74 logits higher; P < 0.001)” should have read (corrected text in bold):

As hypothesized, women not using contraception scored significantly lower on the DAP than those using a method (0.75 logits lower) and those who had not had sex in 30 days (0.74 logits lower; P < 0.001).

REFERENCE