Journal of Digital Imaging
https://doi.org/10.1007/s10278-022-00670-3

Prospective Evaluation of a Machine-Learning Prediction Model for Missed Radiology Appointments

Steven Rothenberg[1,3] · Bill Bame[2] · Ed Herskovitz[1]

Received: 12 November 2021 / Revised: 6 June 2022 / Accepted: 7 June 2022

Abstract

The term "no-show" refers to scheduled appointments that a patient misses, or for which she arrives too late to utilize medical resources. Accurately predicting no-shows creates opportunities to intervene, ensuring that patients receive needed medical resources. A machine-learning (ML) model can accurately identify individuals at high no-show risk, to facilitate strategic and targeted interventions. We used 4,546,104 non-same-day scheduled appointments in our medical system from 1/1/2017 through 1/1/2020 for training data, including 631,386 no-shows. We applied eight ML techniques, which yielded cross-validation AUCs of 0.77–0.93. We then prospectively tested the best performing model, Gradient Boosted Regression Trees, over a 6-week period at a single outpatient location. We observed 123 no-shows. The model accurately identified likely no-show patients retrospectively (AUC 0.93) and prospectively (AUC 0.73, $p < 0.0005$). Individuals in the highest-risk category were three times more likely to no-show than the average of all other patients. No-show prediction modeling based on machine learning has the potential to identify patients for targeted interventions to improve their access to medical resources, reduce waste in the medical system and improve overall operational efficiency. Caution is advised, due to the potential for bias to decrease the quality of service for patients based on race, zip code, and gender.

Introduction

The term "no-show" refers to scheduled appointments that a patient misses or arrives too late to utilize medical resources. No-shows can waste provider time, underutilize limited medical resources, disrupt scheduler workflows, cause a loss of revenue for the radiology department, and most important, deprive patients of important medical tests their physicians ordered. "No-shows" do not include cancellations, modifications, or other situations where prior notice was given.

Accurately predicting no-shows creates opportunities to intervene, ensuring that patients receive needed medical resources. We hypothesized that a machine-learning prediction model can accurately identify individuals at high no-show risk, to facilitate strategic and targeted interventions. The objectives of this study were to train a prediction model using data readily available in our electronic medical record (EMR) and validate this prospectively at a single outpatient-imaging center. If successful, this algorithm could be used to implement interventions to reduce no-shows.

A literature review conducted in 2020 found that 82% of the articles published on predicting missed appointments since 2020 were published during the last 10 years, with logistic regression being the most common algorithm used. Of the 50 studies included in the review, 26% used the same training data and validation data for performance, 62% conducted single validation with split training and validation data, and only 12% of the studies performed a repeat or k-fold validation [1]. Given currently available literature, prospective validation is needed to evaluate model performance.

Methods

This study was determined to be exempt from institutional review board as it was less than minimal risk to human subjects.

✉ Steven Rothenberg
srothenb@uab.edu

[1] Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

[2] University of Maryland Medical System, Baltimore, MD, USA

[3] Department of Radiology, University of Alabama at Birmingham, Birmingham, AL, USA

Published online: 29 June 2022

Model Creation

39 features from our EMR were selected that had the potential for no-show prediction (Table 1). 4,546,104 non-same-day scheduled appointments in our medical system from 1/1/2017 through 1/1/2020 were selected for training data, including 631,386 no-shows. No-shows are recorded in the EMR by technologists after a missed appointment as part of routine practice. We applied eight ML techniques retrospectively, which yielded cross-validation AUCs of 0.77–0.93 (Table 2). Two separate Gradient Boosted Regression Trees models were created using xgboost (https://xgboost.ai/) and catboost (https://catboost.ai/). Xgboost outperformed catboost, therefore the catboost results were dropped. All code was written in python and no formal statistical package was used.

Table 1 List of features and relative contributions from Gradient Boosted Regression Trees model. It is important to note that this model incorporates up to 4-way interactions between features, so although the metrics above estimate overall feature-by-feature contributions, those contributions should not be thought of as "weights" or considered independently of other features. The four metrics therefore summarize or approximate the impact a feature had on the model. See comments on SHAP values in discussion

Description	% Incr	% Decr	Min	Max
Body Mass Index (Most Recent)	0.58	0.42	-5.62	1.22
Appointment Day-of-Year	0.50	0.50	-4.65	1.91
Tobacco User Category	0.65	0.35	-4.22	0.36
Joint Appointment Flag	0.98	0.02	-3.90	1.69
Appointment No Show Count (Previous)	0.29	0.71	-1.21	3.71
Advance Directives	0.43	0.57	-0.49	3.67
Appointment Scheduled From (Epic Module)	0.67	0.33	-2.96	3.56
Appointment Day-of-Week	0.46	0.54	-1.23	3.24
Appointment Department	0.43	0.57	-1.19	3.19
Appointment Department Specialty	0.49	0.51	-2.14	3.07
Appointment LWBS Count (Previous)	0.92	0.08	-2.89	1.69
Provider Type Category	0.40	0.60	-1.97	2.62
Appointment Change Count	0.89	0.11	-2.60	0.88
Referral Flag	0.54	0.46	-0.75	2.45
Appointment Lead Days	0.53	0.47	-1.45	2.36
Appointment Procedure Type	0.51	0.49	-1.26	2.35
Appointment Length	0.57	0.43	-1.64	2.23
Appointment Center (Location)	0.44	0.56	-1.68	2.18
Referral Requested Flag	0.22	0.78	-0.68	2.15
Appointment Normal Status Count (Previous)	0.62	0.38	-2.06	1.55
Zip Code (Patient Permanent Address)	0.42	0.58	-0.85	1.94
Appointment Hour-of-Day	0.48	0.52	-1.68	1.94
Appointment No-Show Ratio	0.71	0.29	-0.78	1.89
Patient Religion Category	0.50	0.50	-1.89	1.00
Appointment Block Category	0.57	0.43	-1.60	1.60
Patient Financial Class	0.31	0.69	-1.11	1.57
Number of Calls (Reminders etc.)	0.77	0.23	-1.51	1.00
Number of Canceled Appointments (Previous)	0.51	0.49	-0.83	1.42
Appointment Confirmation Status	0.59	0.41	-1.35	1.21
Patient Language	0.44	0.56	-1.31	1.01
Patient Ethnic Group	0.25	0.75	-0.71	1.30
Age (on Appointment Date)	0.45	0.55	-1.00	1.26
Homeless Flag	0.00	1.00	-0.08	1.22
Employment Status	0.39	0.61	-0.76	1.21
Veteran Status	0.26	0.74	-0.45	0.91
Marital Status	0.56	0.44	-0.66	0.83
Interpreter Needed Flag	0.68	0.32	-0.80	0.63
Appointment Month	0.50	0.50	-0.52	0.44
Patient Sex	0.44	0.56	-0.23	0.27

Table 2 List of machine learning techniques applied to retrospective data with respective performance as measured by AUC

Machine Learning Technique	AUC
Epic No-Show Model (Logistic Regression)	0.77
Ochsner Model 1 (Logistic Regression)	0.81
Ochsner Model 2 (Neural Network)	0.82
Ridge Regression	0.85
Support Vector Regression	0.88
Random Forrest	0.92
Deep Feedforward Neural Network (i.e. Deep Learning)	0.93
Gradient Boosted Regression Trees	0.93

Table 3 Binned analysis of no-show results based on risk score

Risk Score	# of Shows	# of No Show	Rate of No Show
less than 0.05	535	16	2.9%
0.05–0.10	626	19	2.9%
0.10–0.15	349	20	5.4%
0.15–0.20	219	22	9.1%
0.20–0.25	151	10	6.2%
0.25–0.30	93	6	6.1%
0.30–0.35	78	12	13.3%
0.35–0.45	62	12	16.2%
Above 0.45	28	6	17.6%
Total	2141	123	5.4%

Prospective Validation

The best performing model, Gradient Boosted Regression Trees, was tested prospectively, over a 6-week period by calculating a no-show risk score two weeks prior to every outpatient's scheduled appointment at a single outpatient location. Outcomes for all visits were derived from the electronic medical record (EMR) after the scheduled appointment. We binned risk-scores in 0.05 intervals and used Microsoft Excel (Redmond, WA) to calculate the AUC.

Statistical Analysis

Statistical significance was calculated using a t-test to compare the risk score of the two groups of patients (show/no-show) with Microsoft Excel data analysis package. Subgroup analysis to determine the relative risk comparing groups of participants with a specific risk score bin was calculated using MedCalc's Relative Risk Calculator.

Results

123 no-shows were observed in 2,264 total scheduled exams (5.4%). An ROC analysis yielded AUC of 0.73 ($p < 0.0005$) (Table 3).

Analysis

Sub-group analysis of risk scores above 0.30 demonstrated a 13.3–17.6% rate of no-show. This high-risk subgroup was three times more likely to no-show compared to average. The high-risk subgroup's (risk score above 0.30) relative risk compared to the low-risk subgroup (risk score ≤ 0.1) was 6.08 (95% interval 12.1 to 4.7, $p < 0.0001$).

Features that most increased the no-show risk score were:

- Joint appointment flag
- Appointment left without being seen count
- Appointment normal status count
- Appointment no show ratio
- Interpreter needed flag

Features that most decreased the no-show risk score were:

- Referral requested flag
- Homelessness flag
- Patient ethnic group
- Veteran status
- Appointment no show count (previous)

Features that had the largest effect size for increasing the risk score were:

- Appointment no show count
- Advance directives
- Appointment day of week
- Appointment department specialty
- Appointment lead days

Features that had the largest effect size for decreasing the risk score were:

- BMI
- Appointment day of year
- Tobacco user category
- Joint appointment flag
- Appointment scheduled from

Discussion

Machine learning has the potential to identify patients who are at risk for missing their radiology appointments. Although our risk model achieved statistically significant results (AUC = 0.73, $p < 0.0005$) for prospective prediction, the prospective performance was worse than all of the retrospective models trained (AUC range 0.77–0.93). Our best performing retrospective model was comparable to a top performing retrospective model in the literature (Kurasawa et. al achieved AUC = 0.958 for missed diabetes appointments) [2]. The difference in performance between retrospective and prospective implementation reinforces the critical need for prospective validation of machine-learning models [3].

Several of the machine learning models performed well, including Random Forest, Deep Feedforward Neural Network, and Gradient Boosted Regression Trees. We are not sure why Gradient Boosted Regression Trees performed best on our retrospective cohort. Our strategy was to empirically select the best performing model to test prospectively after exploring several different machine-learning techniques.

The features of the best performing model analyzed in Table 1 were evaluated using SHAP (SHapley Additive exPlanations) [4] values aggregated over all 5 cross-validation sets where:

- % Incr = percent of time this features increases the risk score
- % Decr = percent of time this feature decreases the risk score
- Min/Max = range of contributions for this feature (pre-logistic-transformation)

The patient population studied were outpatients scheduled for radiology examinations at a single imaging center owned by a large academic medical center located in west Baltimore. Similar to other studies in the literature, one of the most important features for determining no-show risk was the appointment no-show ratio and appointment no-show count.

Algorithm actionability is important for implementation of a model into clinical practice. The highest-risk group (risk score > 0.3) had a no-show rate of 15.1% compared to the lowest-risk group (risk score < 0.1) with a no-show rate of 2.9%. Although this group is 5 times more likely to miss an appointment, the low incidence of no-show events may limit return on investment for cost-intensive interventions (e.g. providing ride-sharing services).

It is worth noting that algorithm bias is a concern in radiology [5]. Underlying racial and socioeconomic disparities may be relevant to our results, as our algorithm used features of religion, race, and zip code. Therefore, interventions for high-risk groups such as appointment double booking may lead to worse experiences for disadvantaged patients in the form of longer outpatient imaging wait times. We advise caution to those considering double-booking high-risk no-show patients.

Room for Improvement

We trained our models using appointment data from before the COVID-19 pandemic. Our initial attempts to implement the model were delayed due to prolonged closure of the outpatient-imaging center starting in March of 2020. When we finally evaluated the model, a shift in patient behavior as a result of the pandemic may have contributed to our observed degraded performance relative to retrospective evaluation results. During prospective implementation, our no-show rate decreased relative to the rate from the training data, which may have been due to flexible working conditions for patients with a societal shift to remote work, among other factors. Training on post-pandemic data may improve model performance.

Conclusion

Machine learning can be used to identify patients at risk for missing their radiology appointments. Our model performed worse on prospective than on retrospective data, but results were still statistically significant with respect to no-show prediction. Our results highlight the importance of, and need for, prospective evaluation of machine-learning models before they can be used for clinical decision-making.

Declarations

Competing Interests The authors declare no relevant conflicts of interest.

References

1. Carreras-García D, Delgado-Gómez D, Llorente-Fernández F, Arribas-Gil A. Patient No-Show Prediction: A Systematic Literature Review. Entropy. 2020;22(6):675. https://doi.org/10.3390/e22060675
2. Kurasawa H, Hayashi K, Fujino A, et al. Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients With Diabetes. J Diabetes Sci Technol. 2015;10(3):730-736. https://doi.org/10.1177/1932296815614866
3. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol. 2019;20(3):405-410. https://doi.org/10.3348/kjr.2019.0025
4. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. 2017. https://doi.org/10.48550/arXiv.1705.07874
5. Allen B, Dreyer K. The Role of the ACR Data Science Institute in Advancing Health Equity in Radiology. J Am Coll Radiol JACR. 2019;16(4 Pt B):644-648. https://doi.org/10.1016/j.jacr.2018.12.038