DOI: 10.1111/1475-6773.13200

RESEARCH ARTICLE

HSR Health Services Research

Novel application of approaches to predicting medication adherence using medical claims data

Leah L. Zullig PhD MPH[1,2] | Shelley A. Jazowski MPH[2,3] | Tracy Y. Wang MD MHS MSc[4] | Anne Hellkamp MS[4] | Daniel Wojdyla MS[4] | Laine Thomas PhD[4,5] | Lisa Egbuonu-Davis MD MPH MBA[6] | Anne Beal MD MPH[6] | Hayden B. Bosworth PhD[1,2,7,8,9]

[1]Center of Innovation to Accelerate Discovery and Practice Transformation, Durham Veterans Affairs Health Care System, Durham, North Carolina
[2]Department of Population Health Sciences, Duke University, Durham, North Carolina
[3]Department of Health Policy and Management, University of North Carolina, Chapel Hill, North Carolina
[4]Duke Clinical Research Institute, Duke University, Durham, North Carolina
[5]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina
[6]Global Patient Centered Outcomes and Solutions, Sanofi, New York, New York
[7]School of Nursing, Duke University, Durham, North Carolina
[8]Department of Psychiatry and Behavioral Sciences, Duke University, Durham, North Carolina
[9]Department of Medicine, Duke University, Durham, North Carolina

Correspondence
Hayden B. Bosworth, PhD, Center of Innovation to Accelerate Discovery and Practice Transformation, Durham VA Health Care System and Department of Population Health Sciences, Duke University, 411 W. Chapel Hill Street, Suite 600, Durham, NC 27701, USA.
Email: boswo001@duke.edu

Funding information
Sanofi; VA Health Services Research and Development (HSR&D), Grant/Award Number: CDA 13-025 and VA HSR&D 08-27; Center of Innovation for Health Services Research in Primary Care, Grant/Award Number: CIN 13-410; PhRMA Foundation; Proteus Digital Health; AstraZeneca; Bristol Myers Squibb; Cryolife; Portola; Regeneron

Abstract

Objective: To compare predictive analytic approaches to characterize medication nonadherence and determine under which circumstances each method may be best applied.

Data Sources/Study Setting: Medicare Parts A, B, and D claims from 2007 to 2013.

Study Design: We evaluated three statistical techniques to predict statin adherence (proportion of days covered [PDC ≥ 80 percent]) in the year following discharge: standard logistic regression with backward selection of covariates, least absolute shrinkage and selection operator (LASSO), and random forest. We used the C-index to assess model discrimination and decile plots comparing predicted values to observed event rates to evaluate model performance.

Data Extraction: We identified 11 969 beneficiaries with an acute myocardial infarction (MI)-related admission from 2007 to 2012, who filled a statin prescription at, or shortly after, discharge.

Principal Findings: In all models, prior statin use was the most important predictor of future adherence (OR = 3.65, 95% CI: 3.34-3.98; OR = 3.55). Although the LASSO regression model selected nearly 90 percent of all candidate predictors, all three analytic approaches had moderate discrimination (C-index ranging from 0.664 to 0.673).

Conclusions: Although none of the models emerged as clearly superior, predictive analytics could proactively determine which patients are at risk of nonadherence, thus allowing for timely engagement in adherence-improving interventions.

KEYWORDS
biostatistical methods, chronic disease, medicare

1 | INTRODUCTION

Approximately half of American adults report being prescribed at least one drug in the last 30 days, and nearly one-quarter are prescribed three or more daily medications.[1] Both in the United States and globally, estimates suggest that 50 percent of patients with chronic diseases do not take their medications as prescribed.[2] This lack of adherence leads to potentially avoidable morbidity and mortality, as well as unnecessary health care use and costs.[3] There have been numerous interventions seeking to improve medication adherence, and many of these have been successful in clinical and behavioral trials.[4-8] For these interventions to be most effective, it is important that they target the patients with the greatest need—those who struggle to take their medications as prescribed—as well as identify specific barriers to appropriate medication use. Ideally, clinicians and researchers will identify patients at risk of nonadherence *before* they experience adherence problems. Traditional methods of identifying patients who are nonadherent, or at risk of nonadherence, rely upon either asking patients whether they are having difficulty taking their medications on an individual basis or reviewing pharmacy claims data to determine which patients missed prescription medication refills. These methods tend to be costly, time consuming, and often focused on problems that have already occurred.

Predictive analytics may be a more sustainable, proactive approach for identifying patients at risk of medication nonadherence.[9] Predictive analytics provide an opportunity to segment patient populations based on their established likelihood of nonadherence using predetermined characteristics. These predetermined characteristics can be based on known factors associated with nonadherence, such as minority race, low socioeconomic status, presence of multiple chronic conditions, and/or being prescribed many medications, among other factors.[10] The potential power of predictive analytics is that it enables health care systems to capitalize on limited resources by purposefully and proactively identifying patients at risk of nonadherence and then targeting interventions for those patients.

As a case study to illustrate the novelty of using predictive analytics to support medication adherence improvement, we used Medicare claims data to evaluate what variations may be in common analytical models to predict adherence to lipid-lowering medications. We chose lipid-lowering medications because such therapy is recommended after a myocardial infarction (MI),[11] and dyslipidemia is a common condition that requires chronic and daily medication. In addition, there is a breadth of research informing ways to improve adherence to lipid-lowering medications,[12,13] and there are clear quality indicators to support cholesterol adherence.[14,15]

2 | METHODS

2.1 | Data source and eligibility criteria

We used the fee-for-service Medicare 5 percent sample to identify patients with an acute MI-related admission from 2007 to 2012. The dataset was comprised of Part A (inpatient), Part B (outpatient), and Part D (prescription medication) claims. We restricted our analysis to patients who were discharged to their homes without hospice care, survived at least 1 year after discharge, had Part D coverage for at least 9 months prior to admission and 1 year following discharge, and filled a statin prescription at or within 30 days after discharge. Patients with a prior statin supply had to fill a prescription within 30 days following the end of their previous supply. We considered the first admission for patients with multiple qualifying admissions.

2.2 | Definition of medication adherence

To define medication adherence, we focused on the implementation phase (extent to which actual dosing corresponds to the prescribed regimen from treatment intiation to the last dose) of the Ascertaining Barriers to Compliance (ABC) taxonomy.[16-18] We calculated the proportion of days covered (PDC), which represents the proportion of days in a given period for which the patient had medication supply, for lipid-lowering medications.[19] To calculate PDC, accounting for prior medication supply, we considered patients that had Part D coverage for at least 90 days prior to the beginning of the study period and did not count time spent in an inpatient care setting (eg, acute care, long-term care, skilled nursing facility, etc) in the denominator. We defined adherence as a dichotomous variable, and, consistent with Pharmacy Quality Alliance (PQA) thresholds,[20] we considered patients adherent to lipid-lowering medication if PDC ≥ 80 percent.

2.3 | Statistical analysis

We conducted statistical analysis in SAS version 9.4M5. With the exception of race (missing for 0.2 percent of patients), demographics (age, gender, region, insurance status) were complete in the data. For inclusion in models, we imputed unknown race as white, since the majority of the available sample was white (86 percent). We present categorical variables as percent (count) and continuous variables as median (25th-75th percentiles) unless otherwise noted.

We evaluated and compared three statistical techniques to predict adherence: standard logistic regression with backward selection of covariates, least absolute shrinkage and selection operator (LASSO), and random forest. Our research team selected these methods to represent three usual approaches to prediction:

traditional parametric modeling, shrinkage estimation, and machine learning methods, respectively. Logistic regression with backward selection (alpha to stay = 0.05) is vulnerable to overfitting and is not optimized for the purpose of prediction.[21] LASSO addresses these problems by minimizing Akaike's information criteria (AIC) and shrinking parameters to avoid overfitting.[22,23] Random forest is more flexible than the preceding methods and can identify interactions and nonlinearity that is not prespecified by investigators. Thus, we expected that random forest would perform better if the true model were complex and not summarized by simple main effects.[24,25] We evaluated models for discrimination by the C-index (range 0.5 [noninformative] to 1.0 [perfect prediction])[26] and calibration by decile plots comparing model predictions to observed event rates. Specifically, we calculated these in the full dataset for the logistic regression model; used an 80 percent training dataset and a 20 percent validation dataset in the LASSO regression model; and used 10-fold cross validation for the random forest model.

3 | RESULTS

3.1 | Patient characteristics

There were 757 beneficiaries who had at least one otherwise qualifying admission, were on statin prior to the admission, and still had supply on hand at discharge, but did not refill their statin within 30 days of the end of their previous supply, and so were excluded from analysis. We identified 11 969 patients with hospitalizations for MI between January 1, 2007, and December 27, 2012. On average, patients were white (86 percent) and women (58 percent). Some patients were dually eligible for Medicaid (29 percent) and approximately one-third lived in rural areas (34 percent). Patients were well-distributed geographically, with the largest populations living in the Southern (41 percent) and Midwestern (25 percent) regions. Many patients were diagnosed with hypertension (91 percent), dyslipidemia (84 percent), and diabetes (45 percent). Congestive heart failure, prior MI, and cerebrovascular disease were present in 43 percent, 28 percent, and 26 percent of patients, respectively. Accordingly, many patients had been prescribed a statin (39 percent) in the prior six months and adherence was generally high (median PDC of 97 percent). The proportion of the population whose PDC was ≥ 80 percent in the year following discharge was 64 percent (n = 7642) (median PDC of 89 percent). We present additional information about patient demographics, comorbid conditions, and health care use in Table 1.

3.2 | Logistic regression results

Patients prescribed statins in the prior 6 months had much greater odds of adherence (odds ratio [OR] 3.65, 95% confidence interval [CI] 3.34-3.98) compared to patients who newly filled a statin post-MI. We also identified significant associations between odds of adherence and race ($P < .001$), number of cardiovascular medications

TABLE 1 Baseline and in-hospital characteristics of analysis cohort

	Study population (N = 11 969)
Demographics	
Age median (Q1-Q3)	76 (71-82)
Female %, n	57.6 (6891)
Race %, n	
White	85.7 (10 254)
Black	8.0 (952)
Other	6.4 (763)
Medicaid dual eligible %, n	29.4 (3524)
Resides in rural area %, n	34.2 (4097)
Region %, n	
Northeast	18.9 (2263)
Midwest	25.4 (3045)
South	40.9 (4891)
West	14.6 (1751)
US territory	0.2 (19)
Medical history[a] %, n	
Hypertension	90.7 (10 854)
Diabetes	45.3 (5421)
Dyslipidemia	84.2 (10 078)
Chronic kidney disease	24.0 (2868)
Cerebrovascular disease	25.6 (3065)
Congestive heart failure	42.6 (5097)
Peripheral arterial disease	31.2 (3731)
Prior MI	28.0 (3353)
Prior PCI	3.6 (429)
Prior CABG	0.8 (100)
Health service use in prior 6 mo	
Hospitalizations %, n	
0	80.4 (9626)
1	13.3 (1586)
≥2	6.3 (757)
SNF, IRF, or LTC stay %, n	4.2 (497)
In-hospital	
Cardiogenic shock %, n	2.3 (277)
Cardiac arrest %, n	3.8 (452)
PCI %, n	49.1 (5874)
Drug-eluting stent %, n	30.4 (3640)
CABG %, n	7.8 (933)
Length of stay mean (SD)	5.9 (3.6)
Median (Q1-Q3)	5 (4-7)
Medications in prior 6 mo	
Statin %, n	39.3 (4703)
Prior statin PDC median (Q1-Q3)	96.7 (87.0-100.0)

(Continues)

TABLE 1 (Continued)

	Study population (N = 11 969)
$P2Y_{12}$ inhibitor PDC %, n	11.8 (1410)
Prior $P2Y_{12}$ inhibitor PDC median (Q1-Q3)	96.2 (88.0-100.0)
Discharge medications	
Statin %, n	100 (11 969)
$P2Y_{12}$ inhibitor %, n	69.3 (8292)
Number of cardiovascular meds[b] mean (SD)	3.2 (0.8)
Median (Q1-Q3)	3 (3-4)

Abbreviations: CABG, coronary artery bypass grafting; IRF, inpatient rehabilitation facility; LTC, long-term care; MI, myocardial infarction; PCI, percutaneous coronary intervention; PDC, proportion of days covered; Q, quarter; SD, standard deviation; SNF, skilled nursing facility.
[a]Occurring 1 y prior to admission or during index admission. Prior MI, PCI, and CABG occurring 1 y prior to admission only.
[b]ACE-inhibitor, angiotensin-receptor blocker (ARB), beta block, statin, or $P2Y_{12}$ inhibitor.

prescribed at discharge ($P < .001$), geographic region ($P < .001$), and having a CABG procedure during their index admission ($P < .001$). Patients hospitalized in the prior six months (OR: 0.87, 95% CI: 0.79-0.97) or diagnosed with diabetes (OR: 0.90, 95% CI: 0.83-0.98) had reduced odds of adherence. We present additional results from the backward stepwise selection logistic regression model in Table 2.

3.3 | LASSO regression results

Based on AIC minimization, the LASSO model selected nearly 90 percent of all candidate predictors (25 of 28 variables), with a majority being similar to those selected by the logistic regression model. The magnitude and direction of the associations were also similar between the two models. In the LASSO model, patients prescribed statins in the prior 6 months also had much greater odds of adherence (OR 3.55). People of minority races had lower odds of adherence compared with white patients (black OR 0.61, other races OR 0.79). Patients who had a CABG during their index admission (OR 1.28) and who had ≥4 cardiovascular medications at discharge (compared to ≤2 medications; OR 1.28) also had greater odds of adherence. We present additional results from the LASSO regression model in Table 2.

3.4 | Random forest results

When assessing variable importance, defined as the change in mean squared error (MSE) averaged across all decision trees, statin prescriptions in the prior 6 months, age, and index admission length of stay were the most important when building the model (training subset). We also found that the number of cardiovascular medications prescribed at discharge and geographic region were important variables in developing the model. Similar to the logistic regression and LASSO regression models, statin prescriptions in the prior 6 months were the most important variable in predicting lipid-lowering medication adherence (validation subset).

3.5 | Comparison of models

We examined the performance of the standard logistic regression model with backward selection among patients in the 5 percent Medicare sample; discrimination was moderate, with a C-index of 0.673. There were small differences in the coefficients after re-estimation of the LASSO regression model, indicating similar effects of the predictors on medication adherence. The discrimination of the LASSO regression model remained moderate (C-index 0.677 for training dataset and C-index 0.664 for validation dataset). The discrimination of the third model, the random forest-generated model, was also moderate (C-index 0.666).

4 | DISCUSSION

Nonadherence to prescribed statin therapy remains a problem, even among patients recently hospitalized for an acute MI. We found that patients who newly filled a lipid-lowering medication, or those who are beginning lipid-lowering therapy after a long interruption, are the most likely to experience problems with medication persistence (time period between initation and last dose). [16-18] While some solutions to improve adherence exist,[4-8] we assert that it is critical that interventions be targeted toward patients who would benefit most—those new to therapy who are likely to struggle with taking their medications as prescribed. Predictive analytics has the potential to proactively identify patients who are likely to experience adherence problems, specifically problems with persistence over time, so that they can be engaged in adherence-improving interventions. How well these analytics can identify "pre-" nonadherent patients has significant implications for health care providers, payers, and pharmacy benefit managers, among others, who are tasked with selecting and implementing measures to improve adherence and population health.

Our analysis compared three different types of models—logistic regression, LASSO regression, and random forest-generated models—to determine best practice for predicting medication nonadherence in the context of chronic lipid-lowering medications. We found that all three modeling approaches had moderate discrimination, meaning that the models performed reasonably well. Although the LASSO model selected nearly all candidate predictors (unlike the logistic regression model), no one model emerged as being clearly superior to the others in predicting medication adherence. Relative to the LASSO regression and random forest models, logistic regression is more common in the scientific literature. Additionally, clinicians and policy makers may be more familiar with interpreting odds ratios and *P*-values resulting from logistic regression models rather than coefficients resulting from other modeling techniques (eg, conventional confidence intervals and *P*-values are not available in the

TABLE 2 Model comparisons

	Logistic Reg[a]	LASSO Reg[b]	Random forest
Statin in the prior 6 mo[c]	3.65 (3.34, 3.98)	3.55	
Race			
Black vs White	0.61 (0.52, 0.70)	0.61	
Other vs White	0.78 (0.65, 0.92)	0.79	
Number of cardiovascular meds at discharge			
≥4 vs ≤2	1.42 (1.27, 1.58)	1.28	
3 vs ≤2	1.25 (1.12, 1.40)	1.16	
Region[d]			
Northeast vs South	1.23 (1.10, 1.37)	1.20	
Midwest vs South	1.15 (1.04, 1.27)	1.11	
West vs South	1.04 (0.92, 1.17)	1.00	
CABG during index admission	1.33 (1.15, 1.55)	1.28	
Medicaid dual eligible	1.14 (1.04, 1.25)	1.07	
Any hospitalization in prior 6 mo	0.87 (0.79, 0.97)	0.89	
Diabetes	0.90 (0.83, 0.98)	0.93	
Chronic kidney disease	1.12 (1.01, 1.23)	1.15	
Nonacute institutional care in prior 6 mo	1.27 (1.03, 1.56)	-	
Cerebrovascular disease	0.91 (0.83, 1.00)	0.91	
$P2Y_{12}$ inhibitor at discharge	-	1.01	
Length of stay during index admission[e]	-	1.03	
PCI during index admission	-	1.04	
Resides in rural location	-	1.03	
Peripheral arterial disease	-	0.96	
Age (10-y increase)	-	1.01	
Congestive heart failure	-	0.98	
Prior MI	-	0.97	
SNF, IRF, or LTC stay in prior 6 mo	-	1.17	
$P2Y_{12}$ inhibitor in prior 6 mo	-	1.05	
Female	-	0.99	
Hypertension	-	0.98	
Cardiogenic shock index admission	-	1.04	
Drug-eluting stent placed during index admission	-	1.01	
Prior CABG	-	1.02	
Model C-index	0.673[f]	0.677 (training) 0.664 (validation)	0.666

Abbreviations: CABG, coronary artery bypass grafting; IRF, inpatient rehabilitation facility; LTC, long-term care; MI, myocardial infarction; OR, odds ratio; PCI, percutaneous coronary intervention; SNF, skilled nursing facility.

[a]For logistic regression, odds ratios for statistically significant variables are displayed in order of most to least important (descending χ^2).

[b]The LASSO regression model selected nearly 90% of all candidate predictors (odds ratios shown). Confidence intervals and *P*-values are not available for this model.

[c]For logistic regression, statin in prior 6 mo has $\chi^2 = 842$ and the next largest χ^2, for race, is 48. For all models, statin in the prior 6 mo is defined as being on a statin 6 mo prior to admission and not having discontinued at the time of admission.

[d]US territories were grouped with Southern states.

[e]Length of stay was log-transformed before modeling. The odds ratio shown is the risk increase for a *doubling* of the length of stay.

[f]C-index = 0.676 in model with all candidate variables.

LASSO model to aid in interpretation). In the absence of a clear advantage for other methods, our study suggests that logistic regression has utility for predicting medication adherence.

None of our models had a strong C-statistic. In our review of the scientific literature, relatively few articles describing predictors of adherence reported a C-statistic, making it challenging to determine model fit and to make comparisons across models. Additional work is needed to identify what factors should be included in models to predict adherence. An important next step is to consider additional data sources (eg, electronic health records [EHRs], population-based registry data, provider/pharmacy characteristics, etc) that could improve model fit. It is worth noting that many of these data sources may not be available in all health care settings. For example, in many health care settings, clinical or health care use data and pharmacy refill data are in separate data systems. In instances where these data are accessible and are readily combined, data may not be available in a timely manner at the point of care. When using claims data alone, as in this analysis, it appears that the modeling approaches we tested perform similarly.

Prior studies have assessed the performance of clinical predictive models, finding, as in our study, that machine learning methods performed equivalently to standard regression analyses.[27-29] Although advanced analytic methods and traditional regression models have comparable discrimination, model performance is often influenced by both the size of the cohort under study and the number of events per variable (EPV).[30-33] Evidence suggests that logistic regression models perform better (in terms of accuracy, parsimony and/or discrimination) in smaller datasets with approximately 20-50 EPV, while random forest models perform well with larger sample sizes and achieve sufficient stability when EPV exceeds 200.[30,31,34,35] In addition, the number and type of variables selected as predictors affect model performance. Previous research has demonstrated that random forest models achieve high performance not only as more variables are selected, but also when a large number of continuous variables are used as predictors.[30,34,36,37] Given these considerations, data composition, quality, and completeness should be of the utmost importance when selecting or merging clinical data sources for prediction modeling. A secondary consideration is the availability of qualified personnel to conduct prediction modeling. Traditional regression models, including logistic regression, may require less specialized training than advanced methods, such as random forest models. While this should not be a primary consideration, it may be a practical one for health care systems engaging in predictive analytic work.

While rich clinical data are available in claims, EHRs, and population-based registries, to optimize use, we suggest that predictive analytics be coupled with patient-reported data. With multifaceted data, there is an opportunity to identify patients at risk of nonadherence, inquire about specific challenges and barriers that each patient is facing (eg, transportation barriers to pick up prescriptions, chaotic lifestyle that prohibits keeping clinic appointments, physical difficulties opening pill bottles, not having a caregiver to help remember to take medications, among others), and then intervene to address that unique combination of issues. Predictive tools give a "birds eye view" of who is most likely to be nonadherent based on their characteristics, and patient-reported data can provide actionable information about predictors of adherence. For example, using "big data" health care delivery partners may see that a patient lives in a lower income neighborhood. Using patient-reported data, health care delivery partners could discover that the same patient is struggling with the cost of a particular medication and connect him or her with a copayment assistance program.[38-41] Through that conversation health care delivery partners might also learn that the patient has difficulty interpreting pill bottle instructions for when and how to take their medication, thus health care delivery partners could connect them with more innovative, pictorial medication packaging. The value of these data sources in tandem is paramount.

Identifying patients at risk for nonadherence is critical for several reasons. First, it informs early interventions and identification of patients who might benefit from an adherence intervention. Identifying people at risk could reduce clinical inertia and improve clinical outcomes, including the achievement of therapeutic targets outlined in clinical guidelines.[42] Second, predicting issues of nonadherence, especially those related to access and convenience, could help determine which patients are eligible for long-term medication use. Increasing medication supply per prescription (eg, 90-day supply) has become a common method used by payers to reduce refill requests and address adherence problems.[43,44] Third, predicting adherence problems also has potential reimbursement implications. If health care delivery partners can inform early intervention to improve medication adherence, there may be subsequent improvement in Medicare star ratings and associated bonus payments.[45] Lastly, the ability to predict and address issues with medication adherence may prevent avoidable, high-cost health service utilization (eg, emergency department visits, hospitalizations, etc) and related health care spending (government, payer, and patient).[46,47]

This study had several limitations that are worth noting. First, our dataset was limited to patients who had a previous fill for their medication. As such, we were only able to consider problems at the implementation phase of medication adherence, not at the initiation phase or primary nonadherence. Since the behaviors involved in initiation and implementation are distinct, we assert that this is a minor limitation. Second, our analytic models used administrative claims data that lack clinical, socioeconomic, and behavioral characteristics that may influence medication adherence, as well as information related to reasons for treatment nonadherence (eg, provider determination, medication costs, etc). Our models inherently only included observable characteristics. A number of factors have been associated with medication nonadherence, including elements of a patient's beliefs and the patient-provider relationship that are not observable in these data.[48,49] Future research should focus on linking data sources to provide a comprehensive view of the complex predictors of adherence. Third, use of fee-for-service Medicare claims limits generalizability to other patient populations (eg, younger, uninsured, etc) or payer systems (eg, commercial, Medicaid), as factors associated with lipid-lowering medication adherence could differ across groups; however, the median age of our cohort is 65 years and this covers the majority of patients

suffering from an acute MI.[50] Fourth, we also considered several factors that were measured one year posthospitalization and this approach may further limit generalizability. Fifth, our definition of medication adherence assumes observed medication supply and/or prescription fills equate with actual medication-taking behavior. While this definition may not be entirely accurate, use of PDC and administrative claims are well-documented in the scientific literature.[43,44,51] Lastly, our analytic models did not account for the effect of the coverage gap on predicting adherence. Due to increased cost-sharing, patients who reached the coverage gap, within 1-year following index admission, may have discontinued treatment.

5 | CONCLUSIONS

All three predictive analytic approaches had moderate discrimination, and none of the models tested proved to be superior to the others; however, traditional logistic regression models may be the best approach to inform health care providers because of their relatively straightforward interpretation. Future research using comprehensive or linked datasets is needed to understand which factors best predict lipid-lowering medication adherence. Despite our modest findings, predictive analytics has the potential to identify patients, especially new users, that may experience issues with medication adherence and, thus, allow for timely engagement in adherence-improving interventions.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project was funded by Sanofi. Dr. Zullig is supported by a VA Health Services Research and Development (HSR&D) Career Development Award (CDA 13-025). Dr. Bosworth is supported by a Research Career Scientist Award from VA Health Service Research and Development (VA HSR&D 08-27). This work was also supported by the Center of Innovation for Health Services Research in Primary Care (CIN 13-410) at the Durham VA Medical Center. Dr. Zullig reports research grants to her home institutions from the PhRMA Foundation, Proteus Digital Health and Dr. Wang reports research grants to the Duke Clinical Research Institute from AstraZeneca, Bristol Myers Squibb, Cryolife, Portola and Regeneron, as well as consulting honoraria from AstraZeneca and Sanofi. At the time of this work, Dr. Egbuonu-Davis and Dr. Beal were employed by Sanofi; Dr. Egbuonu-Davis is currently employed at Danaher. Dr. Bosworth reports research grants to his home institutions from Sanofi, Otsuka, Novo Nordisk, Improved Patient Outcomes, Cover My Meds, PhRMA Foundation and Proteus Digital Health, as well as consulting from Sanofi, Abbott and Novartis. Dr. Thomas, Ms. Jazowski, Ms. Hellkamp and Mr. Wojdyla report no conflicts of interest.

ORCID

Hayden B. Bosworth https://orcid.org/0000-0001-6188-9825

REFERENCES

1. Centers for Disease Control and Prevention/National Center for Health Statistics. Therapeutic Drug Use. https://www.cdc.gov/nchs/fastats/drug-use-therapeutic.htm. Accessed August 9, 2018.
2. Brown MT, Bussell JK. Medication adherence: WHO cares? *Mayo Clin Proc*. 2011;86:304-314.
3. Ho PM, Bryson CL, Rumsfeld JS. Medication adherence: its importance in cardiovascular outcomes. *Circulation*. 2009;119:3028-3035.
4. Haynes RB, Ackloo E, Sahota N, McDonald HP, Yao X. Yoa X. Interventions for enhancing medication adherence. *Cochrane Database Syst Res*. 2008;2:CD000011.
5. Nieuwlaat R, Wilczynski N, Navarro T, et al. Interventions for enhancing medication adherence. *Cochrane Database Syst Rev*. 2014;11:CD000011.
6. Adler AJ, Martin N, Mariani J, et al. Mobile phone text messaging to improve medication adherence in secondary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2017;4:CD011851.
7. Fuller RH, Perel P, Navarro-Ruan T, Nieuwlaat R, Haynes RB, Huffman MD. Improving medication adherence in patients with cardiovascular disease: a systematic review. *Heart*. 2018;104:1238-1243.
8. Zullig LL, Ramos K, Bosworth HB. Improving medication adherence in coronary heart disease. *Curr Cardiol Rep*. 2017;19:113.
9. Zullig LL, Blalock DV, Dougherty S, et al. The new landscape of medication adherence improvement: where population health science meets precision medicine. *Patient Prefer Adherence*. 2018;12:1225-1230.
10. Sabaté E. *Adherence to long-term therapies: evidence for action*. Geneva, Switzerland: World Health Organization; 2003.
11. Mercado MG, Smith DK, McConnon ML. Myocardial infarction: management of the subacute period. *Am Fam Physician*. 2013;88:581-588.
12. Schedlbauer A, Davies P, Fahey T. Interventions to improve adherence to lipid lowering medication. *Cochrane Database Syst Rev*. 2010;3:CD004371.
13. van Driel ML, Morledge MD, Ulep R, Shaffer JP, Davies P, Deichmann R. Interventions to improve adherence to lipid-lowering medication. *Cochrane Database Syst Rev*. 2016;12:CD004371.
14. Department of Health and Human Services. Clinical Quality Measures. https://millionhearts.hhs.gov/data-reports/cqm.html. Accessed August 9, 2018.
15. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63:2889-2934.
16. Vrijens B, De Geest S, Hughes DA, et al. A new taxonomy for describing and defining adherence to medications. *Br J Clin Pharmacol*. 2012;73:691-705.
17. De Geest S, Zullig LL, Dunbar-Jacob J, et al. ESPACOMP medication adherence reporting guideline (EMERGE). *Ann Intern Med*. 2018;169:30-35.
18. Helmy R, Zullig LL, Dunbar-Jacob J, et al. ESPACOMP medication adherence reporting guidelines (EMERGE): a reactive-Delphi study protocol. *BMJ Open*. 2017;7:e013496.
19. Centers for Disease Control and Prevention's National Center for Chronic Disease Prevention and Health Promotion. Calculating proportion of days covered (PDC) for antihypertensive and antidiabetic medications: an evaluation guide for grantees; 2015. https://www.cdc.gov/dhdsp/docs/Med-Adherence-Evaluation-Tool.pdf. Accessed August 9, 2018.
20. Lester CA, Mott DA, Chui MA. The influence of a community pharmacy automatic prescription refill program on Medicare Part D adherence metrics. *J Manag Care Spec Pharm*. 2016;22:801-807.

21. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code Biol Med*. 2008;3:17.
22. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, 2nd edn. New York: Springer; 2009:43-94.
23. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Statist Soc B*. 2011;73:273-282.
24. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R* (corr. 7th). New York, NY: Springer; 2017:303-332.
25. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, 2nd edn. New York, NY: Springer; 2009:587-603.
26. Lo-Ciganic W-H, Donohue JM, Jones BL, et al. Trajectories of diabetes medication adherence and hospitalization risk: a retrospective cohort study in a large state Medicaid program. *J Gen Intern Med*. 2016;31:1052-1060.
27. Miller PE, Pawar S, Vaccaro B, et al. Predictive abilities of machine learning techniques may be limited by dataset characteristics: insights from the UNOS database. *J Card Fail*. 2019;25:479-483.
28. Frizzell JD, Liang LI, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol*. 2017;2:204-209.
29. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn KW. Mortality risk prediction in burn injury: comparison of logistic regression with machine learning approaches. *Burns*. 2015;41:925-934.
30. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. *Int J Med Inform*. 2018;116:10-17.
31. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137.
32. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2011;54:774-781.
33. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373-1379.
34. Couronne R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*. 2018;19:270.
35. Kim SY. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics*. 2009;10:147.
36. van der Ploeg T, Steyerberg EW. Feature selection and validated predictive performance I the domain of Legionella pneumophila: a comparative study. *BMC Res Notes*. 2016;9:147.
37. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8:25.
38. Williams J, Steers WN, Ettner SL, Mangione CM, Duru OK. Cost-related nonadherence by medication type among Medicare Part D beneficiaries with diabetes. *Med Care*. 2013;51:193-198.
39. Cutler RL, Fernandez-Llimos F, Frommer M, Benrimoj C, Garcia-Cardenas V. Economic impact of medication non-adherence by disease groups: a systematic review. *BMJ Open*. 2018;8:e016982.
40. Bhuyan SS, Shiyanbola O, Kedia S, et al. Does cost-related medication nonadherence among cardiovascular disease patients vary by gender? Evidence from a nationally representative sample. *Womens Health Issues*. 2017;27:108-115.
41. Whitley HP. Monetary value of prescription assistance program service in rural family medicine clinic. *J Rural Health*. 2011;27:190-195.
42. Patel AA, Kuti EL, Dale KM, Shah SA, White CM, Coleman CI. Effect of medication assistance program on clinical outcomes in patients with diabetes. *Formulary*. 2006;41:518-522.
43. Liberman JN, Girdish C. Recent trends in the dispensing of 90-day-supply prescriptions at retail pharmacies: implications for improved convenience and access. *Am Health Drug Benefits*. 2011;4:95-100.
44. Lauffenburger JC, Franklin JM, Krumme AA, et al. Predicting adherence to chronic disease medication in patients with long-term initial medication fills using indicators of clinical events and health behaviors. *J Manag Care Spec Pharm*. 2018;24:469-477.
45. Centers for Medicare & Medicaid Services. Medicare 2018 Part C & D Star Ratings Technical Notes; 2017. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/2018-Star-Ratings-Technical-Notes-2017_09_06.pdf. Accessed August 9, 2018.
46. Iuga AO, McGuire MJ. Adherence and health care costs. *Risk Manag Healthc Policy*. 2014;7:35-44.
47. Roebuck MC, Liberman JN, Gemmill-Toyama M, Brennan TA. Medication adherence leads to lower health care use and costs despite increased drug spending. *Health Aff (Millwood)*. 2011;30:91-99.
48. Zeber JE, Manias E, Williams AF, et al. A systematic literature review of psychosocial ad behavioral factors associated with initial medication adherence: a report of the ISPOR medication adherence & persistence special interest group. *Value Health*. 2013;16(5):891-900.
49. Hartz A, He T. Why is greater medication adherence associated with better outcomes. *Emerg Themes Epidemiol*. 2013;10(1):1.
50. McNamara RL, Kennedy KF, Cohen DJ, et al. Predicting in-hospital mortality in patients with acute myocardial infarction. *J Am Coll Cardiol*. 2016;68(6):626-635.
51. Iyengar RN, LeFrancois AL, Henderson RR, Rabbitt RM. Medication nonadherence among Medicare beneficiaries with comorbid chronic conditions: influence of pharmacy dispensing channel. *J Manag Care Spec Pharm*. 2016;22:550-560.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zullig LL, Jazowski SA, Wang TY, et al. Novel application of approaches to predicting medication adherence using medical claims data. *Health Serv Res*. 2019;54:1255–1262. https://doi.org/10.1111/1475-6773.13200