ACR Open Rheumatology
Vol. 1, No. 9, November 2019, pp 552–559
DOI 10.1002/acr2.11068

American College of Rheumatology
Empowering Rheumatology Professionals

Supplementing Claims Data with Electronic Medical Records to Improve Estimation and Classification of Rheumatoid Arthritis Disease Activity: A Machine Learning Approach

Candace H. Feldman,[1] Kazuki Yoshida,[1] Chang Xu,[1] Michelle L. Frits,[1] Nancy A. Shadick,[1] Michael E. Weinblatt,[1] Sean E. Connolly,[2] Evo Alemao,[2] and Daniel H. Solomon[1]

Objective. Previous attempts to estimate rheumatoid arthritis (RA) disease activity using claims data only did not yield high performance. We aimed to assess whether supplementing claims data with readily available electronic medical record (EMR) data might result in improvement.

Methods. We used a subset of the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS) that had linked Medicare claims. The disease activity score in 28 joints with C-reactive protein (DAS28-CRP) was considered the gold standard of measure. Variables in the linked Medicare claims, as well as EMR recorded in the preceding one-year period were used as potential explanatory variables. We constructed three models: "Claims-Only," "Claims + Medications," and "Claims + Medications + Labs (laboratory data from EMR). We selected variables via adaptive LASSO. Model performance was measured with adjusted R2 for continuous DAS28-CRP and C-statistics for binary category classification (high/moderate vs low disease activity/remission).

Results. We identified 300 patients with laboratory data and linked Medicare claims. The mean age was 68 years and 80% were female. The mean (SD) DAS28-CRP was 3.6 (1.6) and 51% had high or moderate DAS28-CRP. For the continuous estimation, the adjusted R2 was 0.02 for Claims-Only, 0.09 for Claims + Medications, and 0.18 for Claims + Medications + Labs. The C-statistics for discriminating the binary categories were 0.61 for Claims-Only, 0.68 for Claims + Medications, and 0.76 for Claims + Medications + Labs.

Conclusion. Adding EMR-derived variables to claims-derived variables resulted in modest improvement. Even with EMR variables, we were unable to estimate continuous DAS28-CRP satisfactorily. However, in claims-EMR models, we were able to discriminate between binary categories of disease activity with reasonable accuracy.

INTRODUCTION

The ability to estimate rheumatoid arthritis (RA) disease activity would be a powerful tool for epidemiologic studies that lack direct disease activity measures such as the Disease Activity Score 28-joint counts (DAS28) (1). Currently, despite being recognized as an important factor when examining RA-related outcomes, patterns of medication use, or medication-related toxicities, disease activity is infrequently accounted for in either electronic medical

This study was funded by Bristol-Myers Squibb. Dr. Feldman has received research funding from the Rheumatology Research Foundation and the National Institute of Health (NIH)/National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (1K23 AR071500). Dr. Yoshida received financial support for his doctoral study from Honjo International Scholarship Foundation. Dr. Solomon receives salary support from NIH-K24AR055989.

[1]Candace H. Feldman, MD, MPH, ScD, Kazuki Yoshida MD, ScD, Chang Xu, MS, Michelle L. Frits, BA, Nancy A. Shadick, MD, MPH, Michael E. Weinblatt, MD, Daniel H. Solomon, MD, MPH: Brigham and Women's Hospital, Boston, Massachusetts; [2]Sean E. Connolly, PhD, Evo Alemao, RPh, MS, PhD: Bristol-Myers Squibb, Princeton, New Jersey

Drs. Feldman and Yoshida contributed equally to this work.

Dr. Feldman has received research funding from Pfizer Pharmaceuticals and Bristol-Myers Squibb. Dr. Yoshida received financial support for his doctoral study from Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer, and ASISA). Dr. Shadick has received research funding from Mallinckrodt, Bristol-Myers Squibb, Crescendo Bioscience, and Sanofi/Regeneron. She has also received consulting fees from Bristol-Myers Squibb (under $5,000). Dr. Weinblatt has received research funding from Bristol-Myers Squibb, Crescendo Bioscience, Sanofi/Regeneron and consults for Abbvie, Amgen, Corrona, Glaxo Smith Kline, Horizon, Lilly, Lycera, Merck, Novartis, Pfizer, Roche, Samsung, Scipher and Set Point. Drs. Connolly and Alemao are employees of Bristol-Myers Squibb. Dr. Solomon receives salary support from institutional research grants from Bristol Myers Squibb, Abbvie, Amgen, Pfizer, Genentech, Janssen and Corrona. No other disclosures relevant to this article were reported.

Address correspondence to Candace H. Feldman, MD, MPH, ScD, or Kazuki Yoshida, MD, ScD, Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, 60 Fenwood Road, Boston, MA 02115. E-mail: cfeldman@bwh.harvard.edu or kazukiyoshida@mail.harvard.edu.

Submitted for publication July 10, 2019; accepted in revised form July 16, 2019.

Correction added on September 18, 2019 after first online publication: The Conclusion section of the Abstract has been updated to correct grammatical errors.

SIGNIFICANCE & INNOVATION

- Previous attempts to estimate rheumatoid arthritis (RA) disease activity using claims data only have been unsuccessful.
- We demonstrated that the use of simple electronic medical record (EMR) variables linked to claims data can moderately improve binary classification (high + moderate vs low disease activity + remission).
- Accurate estimation of continuous disease activity score proved to be difficult even with added EMR variables.
- Model-based classification, for example, can be used to examine treatment effect modification by disease activity categories.

record (EMR)-based studies or population-based administrative claims-based studies. These studies often include populations significantly larger than those available in RA-dedicated cohorts and therefore are sufficiently powered to detect relevant but infrequent outcomes, such as medication-related adverse events or cardiovascular events. However, without the ability to account for RA disease activity, it is often challenging to know the degree to which these adverse outcomes are associated with the exposure or whether they are a result of increased RA disease activity.

Prior researchers have demonstrated challenges to developing and validating administrative claims–based algorithms that can accurately estimate rheumatoid arthritis (RA) disease activity (2–4). Data-driven, machine learning tools are increasingly being used to accurately identify RA patients, to phenotype distinct populations, and to develop algorithms to understand comorbidities and adverse outcomes (5,6). To date, only one study has applied machine learning methods to attempt to estimate RA disease activity using administrative claims–based data (2). Similar to prior studies, however, the final models tested showed weak accuracy.

We aimed to use data-driven, machine learning methods to explore alternative strategies to develop algorithms to estimate DAS28 with C-reactive protein (CRP). We combined claims data with readily available electronic medical record (EMR) variables and laboratory values to construct models to estimate DAS28-CRP.

METHODS

Participants. We utilized the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS), a single-center observational cohort of adults (older than 18 years) with prevalent RA cared for at Brigham and Women's Hospital, which is an urban tertiary care teaching hospital (7). Over 1500 patients with confirmed RA by the 1987 ACR criteria (8) have been followed for more than 15 years with annual measurement of disease activity with the DAS28-CRP. Among these patients, we selected individuals with at least 1 year of linked Medicare administrative claims data preceding a disease activity measurement between 2006-2010, the years for which we had existing linked claims. Medicare is the U.S. public insurance for individuals older than 65 years and for a subset of younger individuals with disabilities (9). A subset of BRASS patients with linked Medicare data also had medication benefits through Medicare, known as Medicare Part D, and for these individuals, pharmacy dispensing data were available.

Dependent variables. BRASS recorded DAS28-CRP scores, a version of DAS28 with CRP as an inflammatory marker but without patient global health assessment on an annual basis (10). Each patient potentially had multiple DAS28-CRP measurements, but we focused on the first measurement during the follow-up to avoid correlated dependent variables within each individual, which could complicate the cross-validation process (11). We modeled disease activity in two ways: original continuous form and dichotomized form. We dichotomized the variable as "moderate or high disease activity" (DAS28-CRP at 3.2 or greater) and "low disease activity or possible clinical remission" (DAS28-CRP less than 3.2). We chose these cutoffs based on the treat-to-target strategy for established RA patients (12,13). We will refer to the dichotomized cutoff of 3.2 or greater vs. less than 3.2 as high/moderate vs. low disease activity, respectively, for the purposes of this study, recognizing differing perspectives regarding the definition of a DAS28-CRP cutoff for clinical remission.

Potential explanatory variables. We derived explanatory variables from three sources: Medicare claims data, Medicare Part D pharmacy dispensing data, and EMR data. From Medicare claims, we used ICD-9 codes to identify 26 variables, including demographics, comorbidities, joint replacement surgery, rehabilitation visits, number of RA-related codes, laboratory and imaging use, and health care utilization (See eTable 1 for codes). For a subset of patients, Medicare Part D claims provided medication information regarding biological and conventional disease-modifying antirheumatic drugs (DMARDs), glucocorticoids, and opioids.

We used simple EMR-derived variables collected during routine clinical practice to supplement the claims-derived variables. We did not use variables that may not be available outside of our data sources, such as survey data collected for research purposes from the BRASS cohort. We extracted data via the Research Patient Data Registry (RPDR) (14–16), a centralized clinical data registry consisting of routinely collected data from the EMR. We obtained smoking status, body mass index (BMI), systolic blood pressure, medication use (when Medicare Part D was not available), laboratory abnormalities for RA seropositivity (rheumatoid factor or anticyclic citrullinated peptide), hematocrit, erythrocyte sedimentation rate (ESR), and CRP. When repeated measurements were avail-

able, laboratory abnormality was recorded if any one of the measurements was abnormal. Missing values were handled via the missing category method. For example, a laboratory variable was coded as either normal, abnormal, or missing. We did not pursue natural language processing of EMR free text because we aimed to develop a simple and potentially portable estimation and classification model of disease activity. Tender and swollen joint counts were not available as structured data in our system.

For both claims and EMR data sources, the variable assessment period was the 1-year period preceding the index date on which the first ever DAS28-CRP was recorded in BRASS. This rule was applied to all variables, including relatively stable variables such as seropositivity. For medications, both ongoing therapy and new therapy were considered similarly as "ever use" within this 12-month window.

Modeling strategy. We utilized a form of supervised machine learning, adaptive least absolute shrinkage and selection operator (LASSO). LASSO is a penalized regression that prevents model overfitting by restricting the magnitude of coefficient estimates (regularization) and performs variable selection by setting some coefficient estimates to be zero (17,18). Adaptive LASSO (19) is an improvement upon the original LASSO, which allows a different penalty weight for each coefficient. Our modeling approach involved several steps: 1) initial coefficient estimation with ridge regression, 2) adaptive LASSO for variable selection, and 3) final modeling.

First, we obtained the absolute values of the ridge regression (20) estimates of coefficients. We constructed differential penalties based on the inverse of absolute ridge coefficient estimates (19). These differential penalties ensured that more promising potential explanatory variables were penalized to a lesser extent in the subsequent steps.

Second, we ran an adaptive LASSO for variable selection. The optimal value of the overall penalty term was chosen by minimizing 10-fold cross-validation errors. Importantly, cross-validation results can be dependent on the specific random split of the data when the data set is small. Therefore, we repeated 10-fold cross-validation 10000 times to stabilize this process and to minimize randomness (21,22). We combined these 10000 models by examining the number of times each variable was chosen, and we used variables selected in at least 60% of the adaptive LASSO model fits as the final set of variables, in keeping with prior literature (23).

Third, the final model was fit with multiple regression for DAS28-CRP as a continuous DAS28-CRP, or logistic regression for the dichotomized DAS28 classification (DAS28-CRP 3.2 or greater vs DAS28-CRP less than 3.2). We used adjusted R^2 to compare continuous DAS28-CRP model fits. C-statistics were used to compare the ability of the binary DAS28-CRP classification models to distinguish between high and low disease activity. We additionally calculated sensitivity, specificity, and correct classification rate at the threshold chosen by the Youden index (24) that aims to simultaneously maximize sensitivity and specificity. We used SAS v. 9.4 and R 3.4 [glmnet (25)] for computation.

For the candidate explanatory variables, we considered three increasingly larger potential variable pools to examine how simple EMR variables can improve estimation and classification based on claims variables only: 1) claims only ("Claims-Only"), 2) claims and EMR medications ("Claims + Medications"), and 3) claims, EMR medications, and laboratory values ("Claims + Medications + Laboratory Tests" Model). For EMR variables such as the laboratory test variables, we categorized values into the normal range and the abnormal range and incorporated missing as a category.

RESULTS

Participants and characteristics. We identified 300 adults with RA enrolled in BRASS with 1 year or more of linked Medicare claims preceding their initial DAS28-CRP measurement between 2006-2010. Thirty rheumatologists cared for these 300 patients. The distribution of patient cluster sizes was median 3.5 (interquartile range 1-11). A subset of 95 patients had Medicare Part D medication coverage. Table 1 shows the patient characteristics at the initial DAS28-CRP measurement. The mean age was 68 years, 80% were female, and 92% were white. The mean duration of RA was 21 years, reflecting the nature of the prevalent RA cohort that we were able to link to the Medicare claims. The extent of missingness in EMR data was as follows: BMI 13%, blood pressure (BP) 28%, smoking 12%, rheumatoid factor (RF) 36%, ESR 49%, CRP 19%, and hematocrit (5%). The mean (SD) of DAS28-CRP was 3.6 (1.6). The disease activity categories were as follows:

Table 1. Patient characteristics at the first measurement of DAS28-CRP

Variable	Result
n	300
Age (Mean, SD)	67.94 (9.72)
Gender - female (N, %)	241 (80.33)
Race (N, %)	
white	276 (92.0)
black	14 (4.8)
other	10 (3.3)
DAS28-CRP (Mean, SD)	3.58 (1.62)
DAS28-CRP Category (N, %)	
Clinical Remission	105 (35.0)
Low Disease Activity	41 (13.7)
Moderate Disease Activity	93 (31.0)
High Disease Activity	61 (20.3)
DMARDs (N, %)[a]	
0	92 (30.7%)
1	151 (50.3%)
2	49 (16.3%)
3	8 (2.7%)
RA Disease Duration (Mean, SD)	20.56 (13.28)

Abbreviation: DAS28-CRP, Disease Activity Score in 28 joints with C-reactive protein; DMARD, disease-modifying antirheumatic drug. [a]Number of unique DMARDs prescribed during the 1-year variable ascertainment period.

Table 2. Variable selection and final fit results for continuous disease activity estimation

		Claims Only		Claims + Meds		Claims + Meds + Labs	
Data source	Variable	Selection	Coefficient	Selection	Coefficient	Selection	Coefficient
	(Intercept)	10 000	0.651	10 000	1.029	10 000	0.855
Claims	Number of outpatient visit	0	-	0	-	0	-
	Number of ED visits	0	-	0	-	2	-
	Length of hospitalization	1740	-	2356	-	881	-
	Number of hospitalizations	0	-	0	-	0	-
	Number of chest X-ray	54	-	1589	-	1	-
	Arthrocentesis, yes/no	56	-	1726	-	0	-
	ANA testing, yes/no	0	-	0	-	5	-
	BMD testing, yes/no	0	-	0	-	0	-
	CBC testing, yes/no	54	-	1730	-	5	-
	Anti-CCP testing, yes/no	20	-	420	-	6	-
	Metabolic panel, yes/no	54	-	1738	-	5	-
	HBV/HCV screening, yes/no	1808	-	4204	-	9939	1.655
	Chest CT/MRI, yes/no	0	-	0	-	0	-
	Liver enzymes, yes/no	56	-	2104	-	5872	-
	Tuberculosis tests, yes/no	57	-	1743	-	2693	-
	Age at DAS28-CRP	0	-	0	-	0	-
	CRP, yes/no	9999	−0.948	10 000	−0.944	10 000	−1.298
	RF, yes/no	1398	-	2176	-	4877	-
	ESR, yes/no	1723	-	2270	-	5031	-
	Total Number of ESR/CRP	0	-	0	-	0	-
	Race, Black	465	-	2184	-	104	-
	Race, Non-white/black	6	-	1836	-	1899	-
	Sex	0	-	0	-	0	-
	Charlson Comorbidity Index	0	-	0	-	0	-
	Joint surgeries, yes/no	9999	1.683	9999	1.567	10 000	1.893
	Occupational therapy, yes/no	0	-	0	-	0	-
	Physical therapy, yes/no	62	-	2042	-	787	-
	Total Number of RA Codes	0	-	0	-	0	-
Part D/ EMR	Total number of DMARD use			5442	-	9677	−0.444
	ever use of DMARD			9994	−0.755	9959	−0.117
	ever use of glucocorticoids			3	-	0	-
	ever use of opioids			9937	0.636	6860	0.340
EMR	BMI ≥30					0	-
	25≤ BMI <30					0	-
	BMI <18.5					10 000	13.121
	BMI Missing					9975	0.971
	Systolic BP ≥160					0	-
	Systolic BP 120-159					0	-
	Systolic BP missing					10 000	−1.296
	Smoking, current					1	-
	Smoking, past					302	-
	Smoking, missing					5439	-
	RF abnormal					5547	-
	RF missing					1380	-
	ESR abnormal					8271	0.341
	ESR Missing					13	-
	CRP abnormal					7756	0.588
	CRP missing					5	-
	Hematocrit abnormal					7036	0.340
	Hematocrit missing					10 000	−0.782

Abbreviation: ANA, antinuclear antibody; BMD, bone mineral density; BMI, body mass index; BP, blood pressure; CBC, complete blood count; CCP, cyclic citrullinated peptide; CRP, C-reactive protein; CT, computed tomography; DMARD, disease-modifying antirheumatic drugs; ED, emergency department; EMR, electronic medical record; ESR, erythrocyte sedimentation rate; HBV, hepatitis B virus; HCV, hepatitis C virus; Labs, laboratory test results; Meds, medications; MRI, magnetic resonance imaging; Part D, Medicare Part D prescription claims; RA, rheumatoid arthritis; RF, rheumatoid factor.
All variables, including medications and laboratory results, were as recorded within the 12-month period preceding the DAS28-CRP measurement.

20% in high disease activity (more than 5.1), 31% in moderate disease activity (3.2-5.1), 14% in low disease activity (2.6-3.2), and 35% in clinical remission, here defined as less than 2.6.

Continuous DAS28-CRP estimation. Claims-only data resulted in a highly parsimonious final model with just two binary variables (Table 2). As a result, the proportion of

Table 3. Variable selection and final fit results for binary disease activity classification

		Claims-Only		Claims+ Meds		Claims + Meds + Labs	
Data source	Variable	Selection	Coefficient	Selection	Coefficient	Selection	Coefficient
	(Intercept)	10 000	3.937	10 000	3.737	10 000	2.434
Claims	Number of outpatient visit	0	-	0	-	0	-
	Number of ED visits	5119	-	0	-	0	-
	Length of hospitalization	0	-	3520	-	0	-
	Number of hospitalizations	0	-	0	-	0	-
	Number of chest X-ray	110	-	2	-	0	-
	Arthrocentesis, yes/no	5150	-	3858	-	0	-
	ANA testing, yes/no	636	-	745	-	3	-
	BMD testing, yes/no	0	-	0	-	0	-
	CBC testing, yes/no	118	-	20	-	0	-
	Anti-CCP testing, yes/no	4893	-	4095	-	9755	−0.229
	Metabolic panel, yes/no	5174	-	7205	−0.541	8709	−0.460
	HBV/HCV screening, yes/no	9998	0.557	9999	0.210	10 000	0.571
	Chest CT/MRI, yes/no	0	-	0	-	0	-
	Liver enzymes, yes/no	5139	-	7305	1.306	9588	1.108
	Tuberculosis tests, yes/no	5142	-	7183	0.001	9540	−0.184
	Age at DAS28-CRP	0	-	0	-	0	-
	CRP, yes/no	9997	−0.560	9999	−0.801	10 000	−0.720
	RF, yes/no	6	-	0	-	0	-
	ESR, yes/no	5192	-	7274	−0.228	9583	−0.173
	Total Number of ESR/CRP	0	-	0	-	0	-
	Race, Black	5640	-	8545	0.521	9428	0.307
	Race, nonwhite/black	79	-	21	-	0	-
	Sex	5034	-	5552	-	8961	0.330
	Charlson Comorbidity Index	0	-	0	-	0	-
	Joint surgeries, yes/no	5703	-	7229	0.879	9799	0.729
	Occupational therapy, yes/no	0	-	0	-	0	-
	Physical therapy, yes/no	5147	-	7220	−0.285	9438	−0.217
	Total Number of RA Codes	0	-	0	-	0	-
PartD/ EMR	Total number of DMARD use			7189	−0.430	9501	−0.422
	ever use of DMARD			15	-	0	-
	ever use of glucocorticoids			9999	0.232	10 000	0.216
	ever use of opioids			9996	0.374	9772	0.144
EMR	BMI ≥30			0	-	0	-
	25≤ BMI <30			7189	−0.430	0	-
	BMI <18.5			15	-	0	-
	BMI Missing			9999	0.232	9562	0.450
	Systolic BP ≥160			9996	0.374	0	-
	Systolic BP 120-159					1774	-
	Systolic BP missing					10 000	−0.773
	Smoking, current					0	-
	Smoking, past					0	-
	Smoking, missing					22	-
	RF abnormal					10 000	0.579
	RF missing					7948	0.393
	ESR abnormal					10 000	0.442
	ESR Missing					0	-
	CRP abnormal					9624	0.839
	CRP missing					9586	0.691
	Hematocrit abnormal					7480	0.167
	Hematocrit missing					8793	−0.216

Abbreviation: ANA, antinuclear antibody; BMD, bone mineral density; BMI, body mass index; BP, blood pressure; CBC, complete blood count; CCP, cyclic citrullinated peptide; CRP, C-reactive protein; CT, computed tomography; DMARD, disease-modifying antirheumatic drugs; ED, emergency department; EMR, electronic medical record; ESR, erythrocyte sedimentation rate; HBV, hepatitis B virus; HCV, hepatitis C virus; Labs, laboratory test results; Meds, medications; MRI, magnetic resonance imaging; Part D, Medicare Part D prescription claims; RA, rheumatoid arthritis; RF, rheumatoid factor.
All variables including medications and laboratory results were as recorded within the 12-month period preceding the DAS28-CRP measurement.

continuous DAS28-CRP explained (R^2) was very poor at 0.03 (adjusted R^2 = 0.02). Models derived by an automated variable selection process may not be clinically interpretable. However, in this specific instance, the presence of laboratory tests (ever/never) for viral hepatitis and for CRP remained in the final model. CRP testing may be a surrogate for the need to assess inflammation formally, which is likely due to high disease activity. Viral hepatitis testing may herald the need to switch med-

ications, particularly to biological DMARDs in the setting of higher disease activity or inadequate disease control.

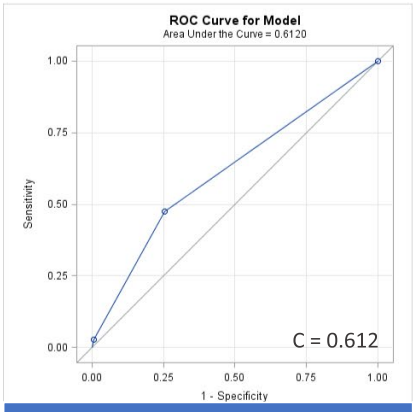
Adding four medication-related variables in the initial candidate variable pool (Claims + Medications) resulted in a much larger final model with 12 variables. The estimation performance was much better, although it still explained a relatively small fraction of continuous DAS28-CRP variability (R^2 = 0.12, adjusted R^2 = 0.09). Viral hepatitis and CRP testing remained in the final model again. Tuberculosis and liver enzyme testing, which may also precede biological DMARDs, were in the final model. Glucocorticoid use and opioid use made it to the final model, but not DMARD use. DMARD use might have been of little value because most patients in this tertiary care center RA cohort were on DMARDs.

Including further EMR variables (Claims + Medications + Laboratory Tests) resulted in a final model with 23 variables. The estimation performance improved further (R^2 = 0.25, adjusted R^2 = 0.18). Laboratory variables (CRP, ESR, RF, and hematocrit) exceeded the model inclusion threshold of 60%. For RF, CRP, and hematocrit, both the abnormal value indicator and the missing indicator remained in the model, meaning whether a measurement was made at all was also informative of the underlying disease activity in addition to the presence of an abnormal measurement.

Binary category classification. Similar to the continuous DAS28-CRP estimation, the binary Claims-only model resulted in a final model with just two variables: the presence of CRP testing and joint surgery (Table 3). The area under the curve (AUC) of the model was 0.61 (Figure 1). At the optimal threshold that maximizes sensitivity and specificity jointly, sensitivity was 47.4% and specificity was 74.7%. Presence of joint surgery may be understood as indicative of more active disease with severe damage.

The inclusion of medication-related variables (Claims + Medications) retained the initial two variables, and ever use of DMARDs and ever use of opioids remained in the model. In comparison to the corresponding continuous model, which retained 12 variables, the binary version only retained only 4 variables. The AUC improved slightly to 0.68 with a sensitivity of 79.2% and specificity of 48.6% at the optimal threshold. Ever use of glucocorticoids was not retained in the final model.

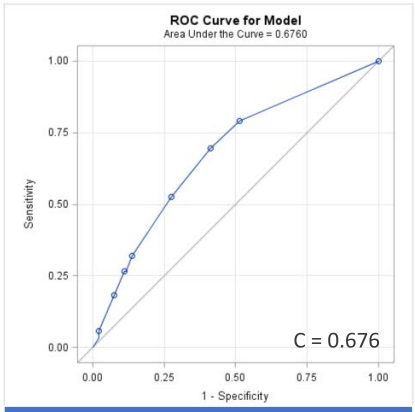
The largest set of candidate variables (Claims + Medications + Laboratory Tests) resulted in a final model with 13 variables, again somewhat smaller than the continuous counterpart with 23 variables. All four variables in the previous model remained. Additionally, testing for viral hepatitis and the total number of DMARDs during the baseline period were included in the final model. From the extended pool of EMR variables, several var-



Claims-Only Model
Cut-off = 0.60

True\Class	REM/LDA	MDA/HDA
REM/LDA	109	37
MDA/HDA	81	73

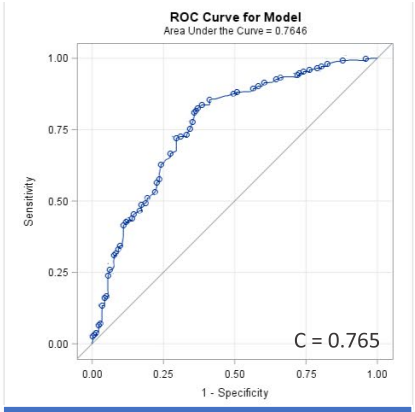
CCR: 61% [54.9, 66.2]
Sensitivity: 47.4% [49.3, 55.6]
Specificity: 74.7% [66.8, 81.5]
PPV: 66.4% [56.7, 75.1]
NPV: 57.4% [50.0, 64.5]
Kappa: 0.22 [0.11, 0.32]



Claims + Meds Model
Cut-off = 0.46

True\Class	REM/LDA	MDA/HDA
REM/LDA	71	75
MDA/HDA	32	122

CCR: 64.3% [58.6, 69.8]
Sensitivity: 79.2% [72.0, 85.3]
Specificity: 48.6% [40.3, 57.0]
PPV: 61.9% [54.8, 68.7]
NPV: 68.9% [59.1, 77.7]
Kappa: 0.28 [0.18, 0.39]



Claims + EMR/Labs Model
Cut-off = 0.43

True\Class	REM/LDA	MDA/HDA
REM/LDA	86	60
MDA/HDA	26	128

CCR: 71.3% [65.9, 76.4]
Sensitivity: 83.1% [76.2, 88.7]
Specificity: 58.9% [50.5, 67.0]
PPV: 68.1% [60.9, 74.7]
NPV: 76.8% [67.9, 84.2]
Kappa: 0.42 [0.32, 0.52]

Figure 1. Performance of binary classification models. Cut-off values indicate the thresholds used to dichotomize estimated probabilities of MDA/HDA into binary classifications (MDA/HDA for ≥ cut-off and REM/LDA for < cut-off). Abbreviation: CCR, correct classification rate; Class, model-based classification; HDA, high disease activity; Labs, laboratory test results; LDA, low disease activity; MDA, medium disease activity; Meds, medications; NPV, negative predictive value; PPV, positive predictive value; REM, clinical remission; True, gold standard.

iables were retained. For certain variables such as BP, only the missingness indicator remained, suggesting that the absence/presence status of having BP recorded was informative enough and actual recordings, when present, did not add much. For ESR and CRP values, respective abnormal value indicators were kept in the final model, grouping normal values and missing category together. The AUC was 0.76 with a sensitivity of 83.1% and specificity of 59.6%.

DISCUSSION

A model that uses readily available data to estimate RA disease activity would be a valuable addition to epidemiologic population-based studies that lack direct disease activity measures. Prior studies have demonstrated significant challenges in developing and validating these algorithms (2,4). We attempted to build on the prior literature by adding EMR variables that should be available from routine practice and are readily extractable from medical records to claims-based data. We also leveraged novel machine learning strategies to allow the data to drive the choice of variables.

In our models, we found that adding EMR-based information modestly improved model performance metrics. However, we were still unable to estimate disease activity in a meaningful way as a continuous measure. Our model that incorporated EMR data, medication and laboratory data, and claims to classify disease activity as a dichotomized measure did result in a reasonable C-statistic (0.76), indicating the ability to distinguish between moderate/high vs. low disease activity with adequate certainty.

These results indicate that addition of simple EMR-derived variables to claims-derived variables could be useful for improving classification of RA disease activity into high and moderate disease activity vs. low disease activity (correct classification rate = 71.3%) but not for accurately estimating its actual numerical values. Importantly, the continuous estimates or binary classification of RA disease activity measure will not add to confounding control if all the covariates are already included in the outcome analysis or propensity score model. However, it may serve as a summary risk score, which can be easier to handle than individual covariates in settings with a limited sample size. A potentially more useful use case of the binary classification is to use it as a stratification variable, for which a correct classification rate of 71.3% may still provide some value. When we are interested in how the effect of a given exposure on the outcome of interest differs by the baseline RA disease activity, stratifying the study cohort by the binary RA disease activity classification may add value beyond what individual covariates can achieve.

We acknowledge several limitations in data and modeling. Ideally, the final model performance should be assessed in a data set completely independent from the entire model building and validation process. However, we did not have a test set because of the small sample size. Although adaptive LASSO is a flexible variable selection strategy, it does not attempt to explore more complex relationships between variables and DAS28-CRP. More advanced supervised machine learning methods, such as deep neural networks (26) and random forest (27), can automatically identify interaction between variables at the cost of being less interpretable and more data-hungry.

Although certain variables included in our final combined model logically correlate with RA disease activity (such as joint surgeries, inflammatory marker elevations, anemia, number and ever use of DMARDs and opioids), our models also incorporated missing EMR data as explanatory variables. For example, presence or absence of EMR recording of systolic BP was deemed more informative than the recorded value itself. Although this is reasonable from a modeling perspective, model interpretability and portability may suffer. The next important step will be determining the degree to which this algorithm may perform in an external cohort with combined EMR-claims data. In such external validation, variables that were important in the BRASS cohort may carry different importance. For example, the tuberculosis screening– and hepatitis screening–related variables contributed to our models. These may have heralded an impending treatment switch or intensification in our local practice and were informative of disease activity. However, this may not generalize to other practice settings. The C-statistic for our Claims + Medications model was only slightly inferior compared to the Claims + Medications + Laboratory Tests model and includes variables that are logically associated with disease activity and are readily accessible in claims (inflammatory markers drawn in the prior year, joint replacement surgery, DMARD use, and opioid use). This may be a reasonable option when the goal is to distinguish moderate/high from low.

In summary, we attempted to improve DAS28-CRP estimation and classification based on claims data by utilizing easily accessible information in the EMR, resulting in a modest improvement for the binary category classification. Numerical estimation of DAS28-CRP was unsatisfactory. External validation of our binary models in a different EMR system is an important future direction. Nonetheless, we believe the present study serves as proof of concept that we can improve our ability to classify RA disease activity in its binary form by supplementing claims data with simple EMR-derived variables.

ACKNOWLEDGMENTS

CHF has received research funding from Pfizer Pharmaceuticals and Bristol-Myers Squibb. KY received financial support for his doctoral study from Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda,

Bayer, and ASISA). CX declared no conflicts. MLF declared no conflicts. NAS has research funding from Mallinckrodt, Bristol-Myers Squibb, Crescendo Bioscience, and Sanofi/Regeneron. She has consulting <5K from BMS. MEW has research funding from and consults for, Bristol-Myers Squibb, Crescendo Bioscience, Sanofi/Regeneron and consults for Abbvie, Amgen, Corrona, Glaxo Smith Kline, Horizon, Lilly, Lycera, Merck, Novartis, Pfizer, Roche, Samsung, Scipher and Set Point. SEC is an employee of Bristol-Myers Squibb. EA is an employee of Bristol-Myers Squibb. DHS receives salary support from institutional research grants from Bristol Myers Squibb, Abbvie, Amgen, Pfizer, Genentech, Janssen and Corrona.

AUTHOR CONTRIBUTIONS

Drs. Feldman and Yoshida had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Feldman, Yoshida, Shadick, Weinblatt, Connolly, Alemao, Solomon.

Acquisition of data. Frits, Shadick, Weinblatt.

Analysis and interpretation of data. Feldman, Yoshida, Xu, Alemao, Solomon.

REFERENCES

1. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
2. Sauer BC, Teng CC, Accortt NA, Burningham Z, Collier D, Trivedi M, et al. Models solely using claims-based administrative data are poor predictors of rheumatoid arthritis disease activity. Arthritis Res Ther 2017;19:86.
3. Ting G, Schneeweiss S, Scranton R, Katz JN, Weinblatt ME, Young M, et al. Development of a health care utilisation data-based index for rheumatoid arthritis severity: a preliminary study. Arthritis Res Ther 2008;10:R95.
4. Desai RJ, Solomon DH, Weinblatt ME, Shadick N, Kim SC. An external validation study reporting poor correlation between the claims-based index for rheumatoid arthritis severity and the disease activity score. Arthritis Res Ther 2015;17:83.
5. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. PLoS One 2015;10:e0136651.
6. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012;19:e162–9.
7. BRASS Study: facts about RA. URL: https://www.brassstudy.org/.
8. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
9. Hoffman ED Jr, Klees BS, Curtis CA. Overview of the Medicare and Medicaid programs. Health Care Financ Rev 2000;22:175–93.
10. Fransen J, Welsing P, De Keijzer RM, Van Riel P. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity. Ann Rheum Dis 2004;62 Suppl:1.
11. Colby E, Bair E. Cross-validation for nonlinear mixed effects models. J Pharmacokinet Pharmacodyn 2013;40:243–52.
12. Smolen JS, Aletaha D, Bijlsma JW, Breedveld FC, Boumpas D, Burmester G, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis 2010;69:631–7.
13. Smolen JS, Breedveld FC, Burmester GR, Bykerk V, Dougados M, Emery P, et al. Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. Ann Rheum Dis 2016;75:3–15.
14. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp 1999;892–6.
15. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. Proc AMIA Symp 2002;552–6.
16. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. AMIA Annu Symp Proc 2006;1044.
17. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1996;58:267–88.
18. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Statist Soc B 2011;73:273–82.
19. Huang J, Ma S, Zhang C-H. Adaptive LASSO For sparse high-dimensional regression models. Statistica Sinica 2008;18:1603–18.
20. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12:55–67.
21. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301–7.
22. Kim J-H. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal 2009;53:3735–45.
23. Austin PC, Tu JV. Bootstrap methods for developing predictive models. Am Stat 2004;58:131–7.
24. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
26. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): The MIT Press; 2016.
27. Breiman L. Random forests. Mach Learn 2001;45:5–32.