Prediction Accuracy With Electronic Medical Records Versus Administrative Claims

Dan Zeltzer, PhD, Ran D. Balicer, MD, PhD,†‡ Tzvi Shir, MA,† Natalie Flaks-Manov, MPH,† Liran Einav, PhD,§|| and Efrat Shadmi, PhD†¶*

Objective: The objective of this study was to evaluate the incremental predictive power of electronic medical record (EMR) data, relative to the information available in more easily accessible and standardized insurance claims data.

Data and Methods: Using both EMR and Claims data, we predicted outcomes for 118,510 patients with 144,966 hospitalizations in 8 hospitals, using widely used prediction models. We use cross-validation to prevent overfitting and tested predictive performance on separate data that were not used for model training.

Main Outcomes: We predict 4 binary outcomes: length of stay (≥ 7 d), death during the index admission, 30-day readmission, and 1-year mortality.

Results: We achieve nearly the same prediction accuracy using both EMR and claims data relative to using claims data alone in predicting 30-day readmissions [area under the receiver operating characteristic curve (AUC): 0.698 vs. 0.711; positive predictive value (PPV) at top 10% of predicted risk: 37.2% vs. 35.7%], and 1-year mortality (AUC: 0.902 vs. 0.912; PPV: 64.6% vs. 57.6%). EMR data, especially from the first 2 days of the index admission, substantially improved prediction of length of stay (AUC: 0.786 vs. 0.837; PPV: 58.9% vs. 55.5%) and inpatient mortality (AUC: 0.897 vs. 0.950; PPV: 24.3% vs. 14.0%). Results were similar for sensitivity, specificity, and negative predictive value across alternative cutoffs and for using alternative types of predictive models.

Conclusion: EMR data are useful in predicting short-term outcomes. However, their incremental value for predicting longer-term outcomes is smaller. Therefore, for interventions that are based on long-term predictions, using more broadly available claims data is equally effective.

Key Words: predictive modeling, electronic medical records readmissions, mortality

(*Med Care* 2019;57: 551–559)

A surge of innovation in digital health is aimed at harnessing predictive modeling and artificial intelligence to predict individual patient outcomes. Such predictions can help improve multiple aspects of medicine,[1–3] including decision support to improve diagnostic accuracy,[4–6] personalized medicine,[7,8] prediction of clinical events and proactive targeting of care,[9–15] and readmission reduction.[16–21] Much attention has been given to data from electronic medical records (EMRs), which contain detailed patient information collected during encounters with providers.

Yet, despite growing efforts to collect and store medical records in electronic form by both hospitals and individual providers, EMR data still present several challenges. Unlike insurance claims data, which are structured to fit a common structure, EMR data vary greatly between facilities, because different vendors and providers use different standards and structures for data encoding and storage. These data are also difficult to share due to both compatibility and privacy concerns. Despite efforts to improve standards, sharing across systems is often limited.[22,23] Consequently, models using EMR data are still mostly trained using data from a single site or system, and applying them widely would require costly retraining and tuning and would heighten privacy concerns due to their sensitive nature.[24,25]

In contrast, although they encode less detailed information on patients, traditional administrative claims data collected for billing purposes have several advantages: they are more structured, standardized across sites, and available for much larger populations (such as the universe of Medicare patients). This gap in the pervasiveness of the 2 data sources raises the question as to the incremental predictive value of EMR data relative to using claims data alone.

Surprisingly, little is known about the additional predictive accuracy gains from using EMR over claims data, despite the large and quickly growing literature dealing with predictive modeling in medicine. Earlier works applied predictive modeling to predict mortality,[26–30] readmissions,[31] length of hospital stay,[11] and other outcomes.[14] However, nearly in all cases, either claims or EMR data were used, but

From the *School of Economics, Tel Aviv University; †Clalit Health Services, Clalit Research Institute, Tel Aviv; ‡Department of Epidemiology, Faculty of Health Sciences, Ben Gurion University, Beersheba, Israel; §Department of Economics, Stanford University, Stanford, CA; ||National Bureau of Economic Research, Cambridge, MA; and ¶Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel.

D.Z. and E.S.: contributed equally.

The data were presented at the Big Data and Precision Medicine 2018 conference at Stanford University.

IRB Reference 0233-16-COM2.

The authors declare no conflict of interest.

Reprints: Dan Zeltzer, PhD, School of Economics, Tel Aviv University, P.O. Box 39040, Tel Aviv 6997801, Israel. E-mail: dzeltzer@tauex.tau.ac.il.

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

ISSN: 0025-7079/19/5707-0551

not both; hence, it is difficult to evaluate the accuracy gains from one over the other. The current study is the first to use EMR and claims data from a single source to compare their power in predicting mortality and hospital readmissions. Earlier work that compared cost predictions using structured diagnosis and medication data from EMR[32] provide evidence that EMR offered only a slight improvement in prediction accuracy over claims data. However, they also acknowledge that future work is required to determine the importance of richer EMR data, such as data on laboratory tests and vital signs.[33]

Motivated by this observation, this study compares the additional accuracy gains of EMR and claims data over claims data alone for predicting readmissions, mortality, and length of stay, using several common prediction methods. We use data from the largest integrated health care delivery system in Israel that provides inpatient care, primary care, and specialty care to >4 million members. These data contain both detailed administrative billing records and EMR and thus allow for a direct comparison of the prediction accuracy of either data type on a large population.

STUDY DATA AND METHODS

Data

We use data from Clalit Health Services, 1 of 4 not-for-profit integrated health care financing and delivery organizations in Israel, which provides Israeli residents with universal tax-funded health care, and therefore continuously collects data on a relatively consistent population. Clalit has over 4 million members (about 54% of the market share). Its patients are admitted to all of Israel's 31 general hospitals, 8 of which it directly owns and operates. It employs >11,000 physicians, 10,000 nurses, and 13,000 administrative staff, and operates over 1500 primary clinics across the country, and provides multiple other outpatient services. Since 2001, it has had a fully integrated EMR IT system used by all providers within the organization. Most patients are continuously covered for years. Therefore, patient data are available over time and across location and care setting, including inpatient and outpatient visits, laboratory and screening tests, imaging, and prescription drugs.

Study Sample

The study sample includes 144,966 index hospital admissions of 118,510 adult patients to 8 hospitals in 2016, their medical history in preceding years, including claims for services billed by external providers, and their outcomes in 2016–2017. We restricted the sample to admissions in Clalit-owned hospitals with an overnight stay, for which Clalit has detailed EMRs. To focus on new episodes of care, the sample also excludes index admission with a prior admission in the previous 30 days. We also excluded admissions of patients who were not continuously covered during the year before the index admission (dropping 2.0% of admissions with discontinuous prior coverage).

A sample of 20,000 patients (Testing Sample) was randomly drawn and kept separate from the remaining sample of 95,510 patients (Training Sample). The test sample was kept intact during the model training stages and used only to test the predictive performance of different models.

Outcomes

Four dichotomous outcomes are predicted, separately: (1) readmission to any hospital in Israel within 30 days of the index admission discharge date, (2) length of stay longer than 7 days, (3) inpatient mortality, defined as death occurring during the index admission, and (4) 1-year all-cause mortality, defined as death within 365 days from the date of admission.

Claims and EMR Variables

Claims records used included patient and encounter data recorded for billing purposes. EMR data used were routinely collected during encounters. The predictors we used are based on data from 2 years before the index admission up to and including the first 2 days of the index admission, except for chronic condition flags, which are based on the entire available patient history. Table 1 describes the classification of variable types to claims and EMR. Note that the same encounters generate both claims and EMR data. For example, a physician visit or a diagnostic laboratory test is recorded on claims along with the associated date, diagnosis codes, and related payments. However, vital signs measured during the office visit (eg, blood pressure) or the exact laboratory test type and result (eg, hemoglobin level) only appear in EMR data.

The information recorded in insurance claims ranges from basic details about patients and their encounters (such as age, gender, or total cost) to detailed coding of diagnoses made during each encounter with providers, from which a more granular picture of the patient morbidity may emerge, for example, by considering the set of chronic conditions a patient was diagnosed with over time. Note that claims, unlike EMR data, are collected by payers and therefore are typically not available to providers in real time. EMRs add yet another layer of information, in the form of indicators that are collected in the diagnostic process.

To understand the relative importance of these different types of information, we defined 4 nested sets of predictors: (1) Basic Claims; (2) Detailed Claims; (3) EMR and Claims combined, which includes data up to but excluding the index admission; and (4) EMR, Claims, and Admission, which include in addition EMR from the first 2 days of the index admission—for assessing the importance of using data collected during the index admission. Basic Claims' variables include patient gender, age in years, and annual utilization counts and cost of medical services, of each of 10 categories: urgent and nonurgent inpatient admissions, emergency room visits, primary care and specialist office visits, outpatient visits, laboratory tests, imaging, prescription drugs, and all other services. Detailed Claims include, in addition to Basic Claims' variables, variables defined on the basis of additional demographic variables, more granular cost and utilization information, and chronic condition flags based on diagnosis codes. Detailed Claims utilization and cost variables are aggregated separately over a period of 30, 60, 90, 180, and 365 days before the index admission. Our detailed

TABLE 1. Claims and EMR Variable Types

Data Category	Basic Claims	Detailed Claims	Claims and EMR	Claims, EMR, and Admission
Demographics				
Age and sex	✓	✓	✓	✓
Additional variables[*]		✓	✓	✓
Cost and utilization				
Aggregate annual summaries[†]	✓	✓	✓	✓
Previous admissions[‡]		✓	✓	✓
Disaggregate summaries[§]		✓	✓	✓
Diagnosis-related covariates				
Chronic conditions' indicators		✓	✓	✓
ACG scores[‖]		✓	✓	✓
Prescription drugs				
Dispensing count, by ATC		✓	✓	✓
Prescription count, by ATC			✓	✓
Laboratory tests, before admission				
Days since last, by type		✓	✓	✓
Results, by type			✓	✓
Outpatient clinical measurements				
BMI			✓	✓
Vital signs (eg, blood pressure)			✓	✓
Substance use			✓	✓
Index admission basic information				
Hospital, ward, urgent/nonurgent		✓	✓	✓
Index admission EMR (first 2 days)				
Diagnoses				✓
Procedures				✓
Vital sign measurements				✓
Laboratory test results				✓
COPS and LAPS[¶]				✓

[*] Additional demographic variables include insurance type, place of residence and service, disability status, socioeconomic status, and immigration status.

[†] Aggregate annual summaries include total cost and number of encounters in each of the following categories: urgent and nonurgent inpatient admissions, emergency room visits, primary care and specialist office visits, outpatient visits, laboratory tests, imaging, prescription drugs, and all other services.

[‡] Days since last previous admission, previous admission duration.

[§] Disaggregate summaries in Detailed Claims are utilization and cost data that are based on 20 service categories and are aggregated separately over the following periods: 30, 60, 90, 180, and 365 days before the index admission.

[‖] Johns Hopkins Adjusted Clinical Groups System predictors.

[¶] See Main text for references and definitions.

ACG indicates Adjusted Clinical Groups; ATC, Anatomical Therapeutic Chemical Classification System; BMI, body mass index; COPS, Comorbidity Point Score; EMR, electronic medical record; LAPS, Laboratory-based Acute Physiology Score.

demographic data contain an indicator for immigration status, which is typically not part of US claims data; however, our results show that it has a negligible effect on the results and could be omitted. We further extracted and summarized information from claims by calculating the Adjusted Clinical Groups (ACG) resource utilization band, a risk measure developed at the Johns Hopkins University and validated and used at Clalit.[34] This system, used by both commercial insurers and noncommercial health care organizations worldwide, depicts overall morbidity on the basis of information from all diagnoses documented at clinical encounter during the previous year. EMR variables include BMI, vital sign measures from encounters, information on drug adherence, and all blood test results. Admission data included EMR variables that were collected during the first 2 days of the index admission, including diagnoses, procedures, vital signs, and other frequently monitored variables, and on-site laboratory tests. We aggregated chronic condition and laboratory test data using Comorbidity Point Score and Laboratory-based Acute Physiology Score, aggregated risk scores that were developed with the aim of real-time prediction of outcomes using admission data.[35,36] We coded missing values as a separate category with its own predictive power (we discretized all continuous variables to turn them into categorical variables). Not imputing missing values is important, because health care records are not missing at random. For example, the set of tests ordered for a patient may provide signals about the patient condition.

Table 2 provides summary statistics for selected predictors. Section B of the Supplementary Digital Content (Supplemental Digital Content 1, http://links.lww.com/MLR/B803) lists the predictors that are included in the data. This initially very large potential set of hundreds of predictors is used to train our models. As discussed in the next section, to avoid overfitting, we use cross-validation methods, so that little to no weight is put on variables that do not in fact contribute to predictive performance.

Predictive Modeling Methods

We use Extreme Gradient Boosting (XGBoost), a sequential ensemble method.[37] This method has 2 appealing properties. First, it involves penalizing the absolute size of the coefficients to avoid overfitting due to the large number of potential predictors. The penalty parameter and other hyperparameters of each model were tuned using 10-fold cross-validation. Second, being an ensemble of classification trees, it

TABLE 2. Descriptive Statistics For Predicted Outcomes and Selected Predictors

	Training Sample	Testing Sample
Sample size		
Patients	98,510	20,000
Admissions	120,483	24,483
Outcomes		
Admission lasts at least 7 d (%)	17.1	17.4
Inpatient mortality (%)	2.6	2.6
30-day hospital readmission (%)	14.6	14.9
1-year all-cause mortality (%)	12.6	12.3
Demographics		
Age (mean) (minimum = 18) (y)	60.5	60.8
Sex (% female)	53.1	52.9
Ethnicity (% Arabs)	19.4	18.9
Supplemental insurance (%)	77.2	77.3
Disability (%)	16.5	16.4
Chronic conditions (%)		
Hyperlipidemia	65.9	66.4
Hypertension	53.9	54.5
Arthropathy	39.4	40.3
Diabetes	32.8	33.2
IHD	30.7	31
Malignancy	22.7	22.9
Arrhythmia	22.5	21.9
Neurological	21.2	21.4
Kidney	19.1	19.9
Gastritis	17.7	17.6
CRF	17.7	17.6
Osteoporosis	15.4	15.7
CVA	15.4	16
Depression	15	15.4
Valvular cardiac	15	15
CHF	14.3	14.2
Prior utilization, mean 1 y count (% nonzero)		
Prescription drugs	72.9 (97.2)	72.9 (97.2)
Laboratory tests	48.2 (91.4)	47.8 (91.5)
Imaging events	3.4 (74.5)	3.4 (74.6)
Ambulatory encounters	7.1 (63.6)	7.1 (63.9)
Emergency room visits	0.9 (45.4)	0.9 (45.7)
Hospital visits	0.9 (39.5)	0.9 (39.8)
ACG score		
Healthy or low	15.2	15
Moderate	46.8	47.3
High or very high	36.9	36.4
Clinical measurements, last measurement, and mean (% nonmissing)		
BMI	27.7 (99.3)	27.7 (99.5)
Diastolic blood pressure (mm Hg)	72.8 (99.2)	72.9 (99.4)
Systolic blood pressure (mm Hg)	124.9 (99.2)	125.2 (99.4)
Hemoglobin (g/dL)	12.7 (96.5)	12.8 (96.5)
Hematocrit (%)	38.9 (96.5)	39 (96.5)
Platelets (1000/μL)	241 (96.5)	241.1 (96.5)
Measurements on the first 2 days of index admission, mean (% nonmissing)		
LAPS	32.7 (89)	32.7 (89.1)
Albumin, first measurement on admission	3.8 (56.6)	3.8 (56.5)
Arterial pH (pCO_2), first measurement on admission	46.5 (21)	46.5 (20.7)
Arterial pH (pO_2), first measurement on admission	48.5 (21)	47.7 (20.7)
CRP, first measurement on admission	26.7 (32)	25 (32.3)
Heart rate, last measurement on first 2 d of admission	78.7 (87.3)	77.3 (87.1)
Blood pressure (systolic), last measurement on first 2 d of admission	124.8 (87.2)	124.8 (87.1)
Blood creatinine—No. measurements on first 2 d of admission	1.6 (99.9)	1.6 (100)

ACG indicates Adjusted Clinical Groups; BMI, body mass index; CHF, congestive heart failure; CRF, chronic renal failure; CRP, C-reactive protein; CVA, cerebrovascular accident; IHD, ischemic heart disease; LAPS, Laboratory-based Acute Physiology Score; pCO_2, partial pressure of carbon dioxide; pO_2, partial pressure of oxygen.

handles well interactions between variables, nonlinear relationships, and missing values (Section C, Supplemental Digital Content 1, http://links.lww.com/MLR/B803). XGBoost can fit arbitrarily well any differentiable criterion function. To check the robustness of the results to the choice of modeling method, sensitivity analyses were performed in which the XGBoost model was replaced with alternative prediction models, including Lasso, ridge regression, and a mixture of both. The analysis was also repeated for the subsample of patients aged 65 or older (for comparability with the age range for which administrative claims from Medicare are available in the United States).

Model Evaluation and Statistical Analysis

The model performance was evaluated using the following statistics: sensitivity, specificity, positive, and negative predictive values (PPV and NPV henceforth), and the number needed to evaluate (which equals 1/PPV). To focus the comparison on the most decision-relevant range of probability threshold, we compared accuracy statistics for each outcome over the deciles of predicted probability scores for this outcome, as estimated by the model using both claims and EMR. For each combination of model, subsample, and set of predictors, the accuracy of predictions was also evaluated using the area under the receiver operating characteristic curve (AUC), a measure of overall model performance. For AUC, 2 measures of relative predictive performance were calculated: First, the differences in AUC between different sets of predictors and second, the percentage of total AUC increase, relative to the baseline of Basic Claims' predictors, that is contributed by each predictor sets. For example, suppose the AUC is 0.7 for a model based on Basic Claims data (our baseline), 0.85 for a model based on Detailed Claims data, 0.87 for a model based on Detailed Claims and EMR data, and 0.9 for a model based on Claims, EMR, and Admission data. In this case, our measure attributes to Detailed Claims three quarters of the overall improvement from using EMR and Detailed claims predictors, relative to the baseline of Basic Claims, $\frac{0.85-0.7}{0.9-0.7}=75\,\%$; EMR accounts for the remaining 25% (including EMR from the index admission, which accounts for $\frac{0.9-0.87}{0.9-0.7}=15\,\%$ of the total improvement). AUC point estimates were obtained using the test sample that was kept separate and not used in training or tuning the models. Because AUC is tested using fresh data, restricting the set of predictors (eg, from both Claims and EMR to Claims alone) may result in *higher* AUC if the reduction in prediction variance is greater than the increase in bias. Confidence intervals (CIs) and significance of the difference between estimated AUC values were obtained using 10,000 bootstrap samples obtained by resampling the test sample with replacement.

To see which predictors contributed the most to prediction accuracy, we also present for each model the 10 predictors with the highest gain. Gain is a measure of the improvement in accuracy brought by a predictor to the classification-tree branches it is on. It is the sum of the reductions in the criterion function following each split that involves the predictor.[37] Gain is normalized to have a sum of 1 over all selected predictors in each model.

STUDY RESULTS

Of the 144,966 index hospital admissions of continuously covered adults in 2016, 17.1% lasted ≥ 7 days, 2.6% ended in death, 14.7% were followed by another admission within 30 days, and 12.5% resulted in death within a year or less.

The estimated AUC of the XGBoost model predicting each of these outcomes trained with each of 4 sets of predictors—Basic Claim; Detailed Claims; Claims and EMR combined; and Claims, EMR, and Admission—is described in Table 3. Estimated AUC values are shown in Figure 1. Calibration is shown in the Supplementary Digital Content (Fig. S1, Supplemental Digital Content l, http://links.lww.com/MLR/B803). The top 10 predictors with the highest gain for each outcome and set of predictors are listed in the Supplementary Digital Content (Tables S8–S23, Supplemental Digital Content l, http://links.lww.com/MLR/B803).

Readmission and 1-Year Mortality

Predicting 30-day readmission, the AUC using detailed information from claims (Detailed Claims) is 0.698 (95% CI: 0.689, 0.707). Using Claims and EMR but excluding index admission data, the AUC is 0.700 (95% CI: 0.691, 0.708); with data from the index admission, it is 0.711 (95% CI: 0.702, 0.720). Relative to the benchmark model using only basic demographic information, cost, and utilization (Basic Claims, AUC: 0.629; 95% CI: 0.619, 0.639), EMR and claims combined provide an increase of 0.082 in AUC ($P < 0.001$). Detailed Claims alone provides an improvement of 0.069 ($P < 0.001$). That is, 84% of the improved prediction accuracy over this baseline that is achieved from using detailed claims and EMR variables can be achieved using claims alone. Restricting the sample to patients aged 65 and older (see Table 3) or to patients with each one of the top most common chronic conditions in our sample (Table S1, Supplemental Digital Content l, http://links.lww.com/MLR/B803) resulted in lower baseline AUC (as expected from conditioning on a significant predictor of the outcome), but in comparable accuracy gains from using EMR in addition to claims in prediction. For predicting 1-year mortality, AUC is 0.902 (95% CI: 0.897, 0.907) using Detailed Claims; 0.912 (95% CI: 0.908, 0.917) using Claims and EMR combined; and 0.921 (95% CI: 0.917, 0.926) using additional admission data [an improvement of 0.009 ($P < 0.001$) relative to pre-admission data). Relative to the Basic Claims' model (AUC: 0.863; 95% CI: 0.857, 0.869], 66.1% of the accuracy gains of 0.058 ($P < 0.001$) from using EMR and claims combined can be realized using detailed claims alone. Data collected during the index admission separately account for 15.5%.

Index Admission Length of Stay and Death

Predicting inpatient mortality using detailed claims, EMR, and admission data combined, the AUC is 0.950 (95% CI: 0.944, 0.957). Relative to the benchmark model (Basic Claims, AUC: 0.841; 95% CI: 0.829, 0.854), 50.7% of the improved accuracy is achieved using detailed claims alone (AUC: 0.897; 95% CI: 0.887, 0.906); admission data account for an additional 47.1%. Predicting whether admissions would last ≥ 7 days, the AUC is 0.785 (95% CI: 0.777, 0.792) using detailed information from claims (Detailed Claims). Using Claims and EMR predictors combined, the AUC is 0.786 (95% CI: 0.779, 0.794) without data from the index admission and 0.837 (95% CI: 0.830, 0.843) with such data. Detailed claims account for 69.4% of the total improved

TABLE 3. Prediction Accuracy of Different Outcomes Using Different Predictor Sets

					Differences in AUC							
	AUC				Detailed Claims—Basic Claims		EMR and Claims—Detailed Claims		EMR, Claims and Admission—EMR and Claims		EMR, Claims and Admission—Basic Claims
	Basic Claims	Detailed Claims	Claims and EMR	Claims, EMR, and Admission	Difference (*P*)	Total (%)	Difference (*P*)	Total (%)	Difference (*P*)	Total (%)	Difference (= Total, 100%) (*P*)
Admission lasts at least 7 d											
Full	0.667	0.785	0.786	0.837	0.118 (< 0.001)	69.4	0.001 (0.056)	0.843	0.051 (< 0.001)	29.8	0.170 (< 0.001)
65+	0.603	0.751	0.749	0.808	0.148 (< 0.001)	72.2	–0.002 (0.868)	−1.13	0.059 (< 0.001)	28.9	0.205 (< 0.001)
Inpatient mortality											
Full	0.841	0.897	0.909	0.950	0.055 (< 0.001)	50.7	0.013 (< 0.001)	11.5	0.041 (< 0.001)	37.8	0.109 (< 0.001)
65+	0.761	0.831	0.849	0.925	0.071 (< 0.001)	43.1	0.018 (< 0.001)	10.9	0.076 (< 0.001)	46.0	0.164 (< 0.001)
30-day hospital readmission											
Full	0.629	0.698	0.700	0.711	0.069 (< 0.001)	83.7	0.002 (0.016)	2.26	0.012 (< 0.001)	14.0	0.082 (< 0.001)
65+	0.610	0.660	0.662	0.674	0.049 (< 0.001)	77.2	0.002 (0.125)	3.42	0.012 (< 0.001)	19.4	0.064 (< 0.001)
1-year all-cause mortality											
Full	0.863	0.902	0.912	0.921	0.038 (< 0.001)	66.1	0.010 (< 0.001)	18.0	0.009 (< 0.001)	16.0	0.058 (< 0.001)
65+	0.776	0.839	0.854	0.869	0.064 (< 0.001)	68.5	0.014 (< 0.001)	15.5	0.015 (< 0.001)	16.0	0.093 (< 0.001)

The percent of total AUC increase is the difference in AUC between 2 models divided by the difference in AUC, shown in the rightmost column, between EMR, Claims and Admission and Basic Claims' model.

Source: Authors' analysis of data from Clalit Health Services.

Full refers to the full sample; 65+ refers to the sample of patients who were 65 years or older when admitted.

AUC indicates area under the receiver operating characteristic curve; EMR, electronic medical record.

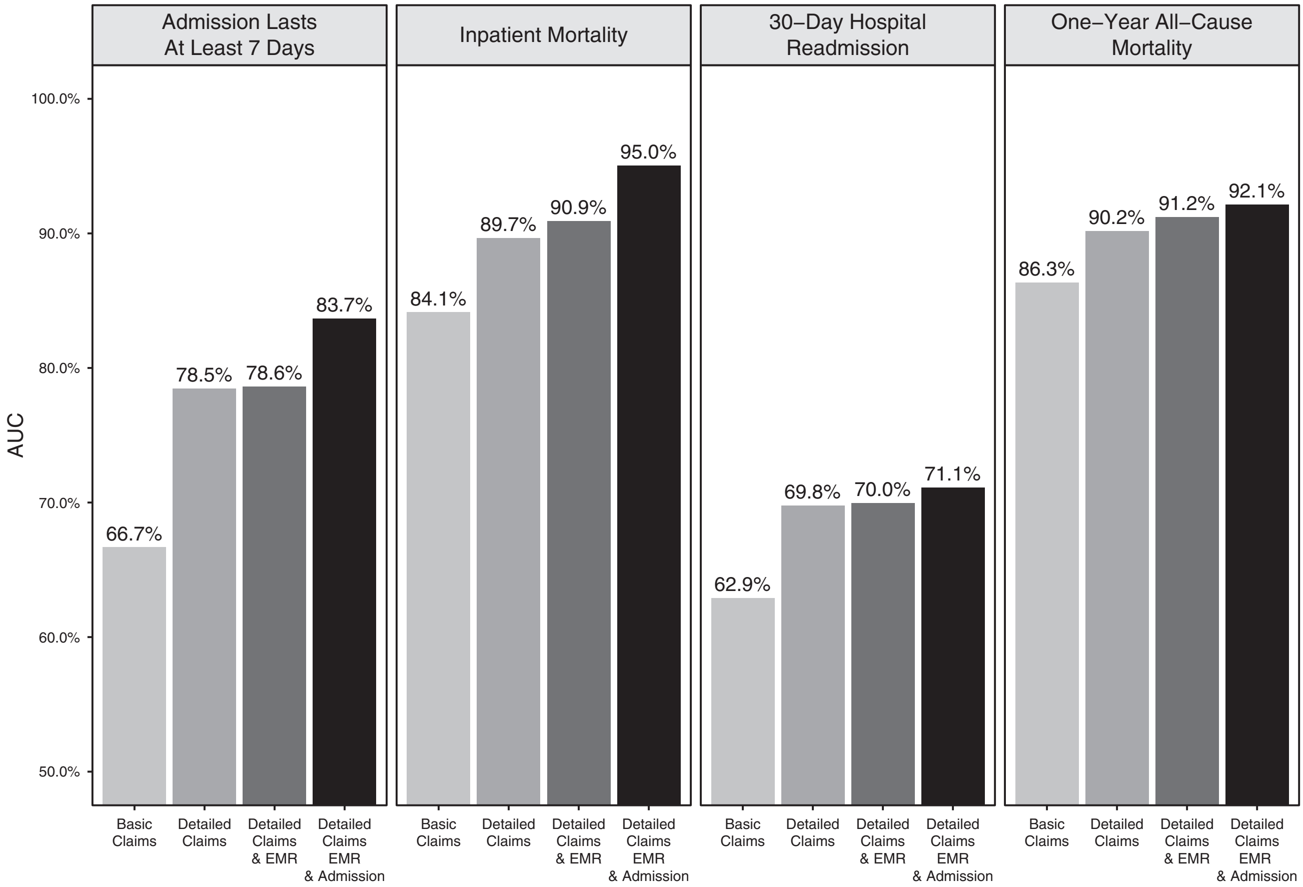

FIGURE 1. Predictive performance by outcome and data type. AUC indicates area under the receiver operating characteristic curve; EMR, electronic medical record.

accuracy over the benchmark model (Basic Claims, AUC: 0.667; 95% CI: 0.658, 0.675) that is achieved using detailed claims and EMR variables. Admission data account for an additional 30%.

Sensitivity Analysis and Alternative Measures of Accuracy

Sensitivity, specificity, and PPV and NPV in predicting each outcome using each of the 3 sets of predictors are shown in Figure 2 (and detailed in Tables S4–S7, Supplemental Digital Content 1, http://links.lww.com/MLR/B803). Each facet shows the result of this analysis for a different accuracy measure and outcome (eg, the top left facet shows the sensitivity of the model in predicting 30-day hospital readmissions). To capture the decision-relevant range of probability thresholds for each outcome, we used the deciles of predicted probability from the empirical distribution of predictors obtained from the (richest) Claims and EMR model. Therefore, the various cut points shown in Figure 2 identify the highest 10%, 20%, and 30% of patients who are most likely to experience the outcome, which corresponds to how predictive models are often used in clinical practice. Deciles were calculated separately for each outcome [eg, the 10% and 90% cutoffs, corresponding to patients with the highest and lowest predicted probability of a 30-day readmission, are 0.258 and 0.056, respectively; for inpatient mortality, the cutoffs are 0.060 and 0.001, respectively, reflecting the different likelihood of these 2 outcomes in the population (Fig. 2); data are available in the Section A.3, Supplemental Digital Content 1, http://links.lww.com/MLR/B803]. For nearly all outcomes and accuracy measures, the relative improvement in performance from using different sets of predictors is similar to that measured using AUC. One exception is the prediction of 30-day hospital readmissions, which for the high specificity range, Basic Claims, exhibits higher specificity (but lower sensitivity) than other models and for the low sensitivity range it exhibits higher sensitivity (but lower specificity). Such results are plausible, as we did not train the models to target specificity and sensitivity directly, but rather to maximize AUC. Additional sensitivity analysis is shown in Section A of the Supplementary Digital Content (Supplemental Digital Content 1, http://links.lww.com/MLR/B803), showing similar results that were obtained when other statistics were used: Net Reclassification Improvement (NRI; Table S3, Supplemental Digital Content 1, http://links.lww.com/MLR/B803) or Number Needed to Evaluate (NNE; Fig. S2, Supplemental Digital Content 1, http://links.lww.com/MLR/B803) were used to measure improvement in accuracy, and when other models were used instead of XGBoost (Table S2, Supplemental Digital Content 1, http://links.lww.com/MLR/B803).

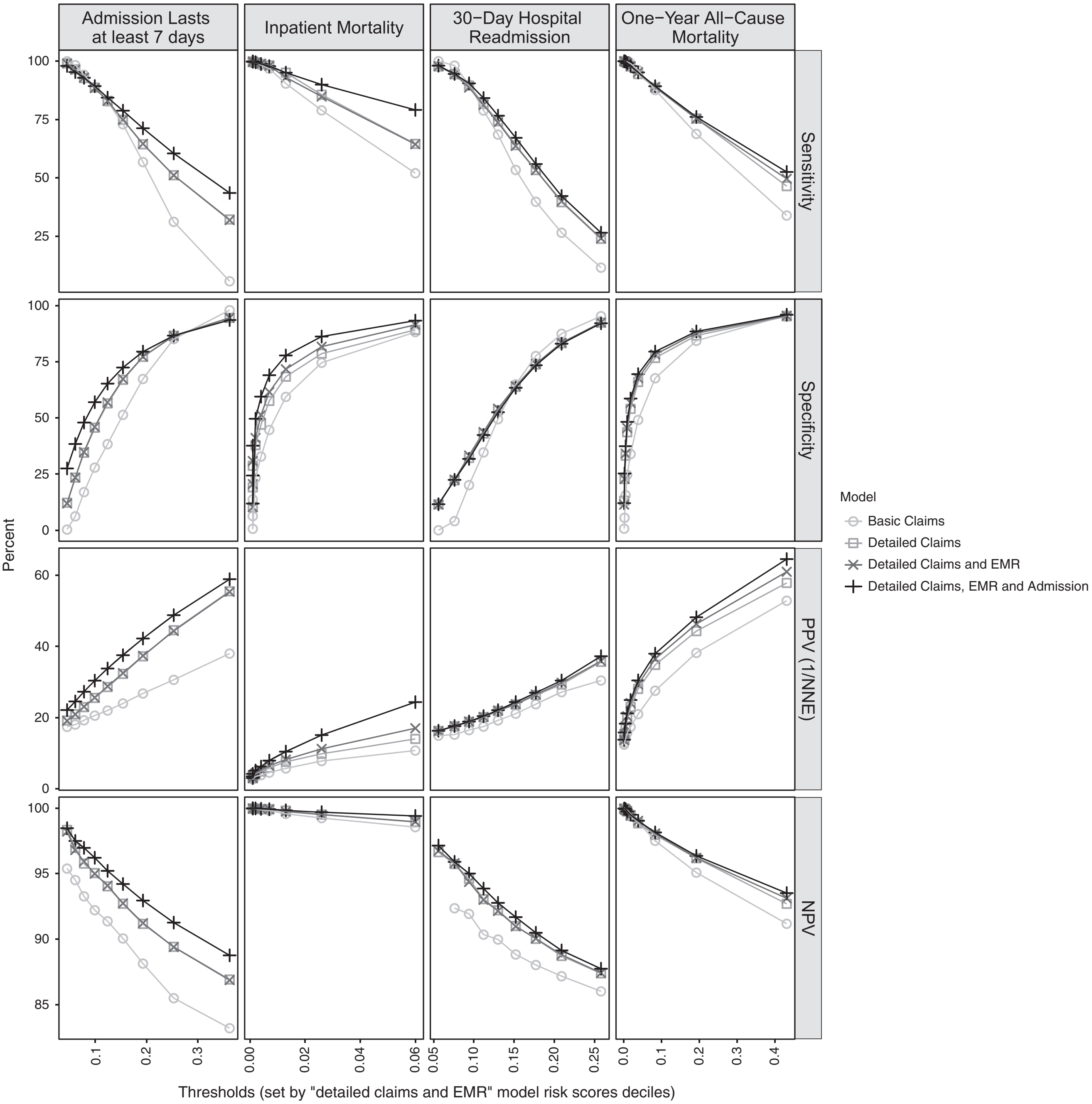

FIGURE 2. Sensitivity, specificity, NPV, and PPV by outcome and data type. Each facet shows the result of 1 analysis of predictive accuracy for a given measure and outcome. Rows correspond to different accuracy measures. Columns correspond to different outcomes. The cutoffs of predicted probability are based on the deciles of the empirical distribution of the predicted probability obtained of the (richest) Claims, EMR, and Admission model. That is, the rightmost vertically aligned data point in each facet corresponds to the threshold identifying the top 10% of patients most likely to experience the outcome, the second to the left to the top 20%, etc, all the way to the bottom 10% for the leftmost points. Deciles were calculated separately for each outcome. *Source*: Authors' calculations using Clalit Health Services data. The data used to generate this figure are tabulated in the Supplementary Digital Content (Supplemental Digital Content l, http://links.lww.com/MLR/B803). EMR indicates electronic medical record; NNE, number needed to evaluate; NPV, negative predictive value; PPV, positive predictive value.

CONCLUSIONS

Predictive modeling is increasingly driving innovation in health care. But if efforts are premised on using EMR data, advances in prediction technology may be impeded by lack of data access and standardization. Using both EMR and data generated from routine encounters within an integrated insurance and health care delivery system in Israel covering 4 million lives, this study evaluated the gains in prediction accuracy from using EMR in addition to more widely available claims data in predicting several common outcomes: hospital readmission, length of stay, and mortality. Findings show that EMR predictors substantially improve accuracy in predicting admission length of stay and inpatient mortality. These results are similar to others that have shown that in-hospital data may significantly contribute to prediction accuracy.[38,39] However, for longer-term outcomes, 30-day readmission and 1-year mortality, most of the gains in predictive accuracy, when combined EMR and claims data are used, can be achieved using claims data alone, given that claims data are detailed enough and include similar information to the data presented here.

These results suggest that that gains from EMR data may depend on the outcome and prediction horizon. EMR data may be particularly important where rich EMR data exist during the relevant time frame for prediction, such as in predicting clinical events occurring during a hospital admission.[40] EMR data may also be important when claims data are of lower quality, when elements such as unstructured text and image data are uniquely important, or when the real-time prediction is critical. However, for longer-term outcomes, claims data, particularly more granular information such as diagnoses codes, may be sufficient for predictive modeling. Therefore, results do suggest that for a variety of applications focused on prediction, the availability of EMR is not a necessary condition for accurate predictive modeling, and that nearly as high an accuracy can be obtained using detailed, yet standard claims, that are already available for a larger population.

One limitation is that our data come from 1 health care system. However, our prediction performance is comparable to previous predictions of readmission or mortality outcomes in other settings.[38–41]

In sum, with the advent of EMR, health care organizations continuously strive for "meaningful use" to support point-of-care decisions, such as those related to the identification of patients at high risk for a range of adverse hospitalization outcomes.[42] Yet, this study shows that gains from EMR vary by context. The predicted outcome, the prediction horizon, and the accuracy and completeness of EMRs and Claims databases should all be considered in decisions on the types of prediction models used for high-risk case identification.

REFERENCES

1. Yoo I, Alafaireet P, Marinov K, et al. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*. 2012;36:2431–2448.
2. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff*. 2014;33:1163–1170.
3. Pencina MJ, Peterson ED. Moving from clinical trials to precision medicine: the role for predictive modeling. *JAMA*. 2016;315:1713–1714.
4. Tang PC, Ralston M, Arrigotti MF, et al. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc*. 2007;14:10–15.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
6. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. *PLoS ONE*. 2017;12:e0177726.
7. Weiss JC, Natarajan S, Peissig PL, et al. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. *AI Mag*. 2012;33:33–45.
8. De Castro DG, Clarke PA, Al-Lazikani B, et al. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. *Clin Pharmacol Ther*. 2013;93:252–259.
9. Phillips JL, Rolley JX, Davidson PM. Developing targeted health service interventions using the PRECEDE-PROCEED model: two Australian case studies. *Nurs Res Pract*. 2012;2012:279431.
10. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff*. 2014;33:1123–1131.
11. Cai X, Perez-Concha O, Coiera E, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J Am Med Inform Assoc*. 2015;23:553–561.
12. Choi E, Bahadori MT, Schuetz A, et al. Doctor AI: predicting clinical events via recurrent neural networks. *JMLR Workshop Conf Proc*. 2016;56:301–318.
13. Aczon M, Ledbetter D, Ho L, et al. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. arXiv preprint arXiv:1701.06675; 2017.
14. Avati A, Jung K, Harman S, et al. Improving palliative care with deep learning: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), November 13–16, 2017. Kansas City, MO: IEEE; 2017:311–316.
15. David G, Smith-McLallen A, Ukert B. The effect of predictive analytics on healthcare utilization. *J Health Econ*. 2018;64:68–79.
16. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306:1688–1698.
17. Donzé J, Aujesky D, Williams D, et al. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern Med*. 2013;173:632–638.
18. Betihavas V, Frost SA, Newton PJ, et al. An absolute risk prediction model to determine unplanned cardiovascular readmissions for adults with chronic heart failure. *Heart Lung Circ*. 2015;24:1068–1073.
19. Shadmi E, Flaks-Manov N, Hoshen M, et al. Predicting 30-day readmissions with preadmission electronic health record data. *Med Care*. 2015;53:283–289.
20. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10–13, 2015. NSW, Australia; 2015:1721–1730.
21. Zhou H, Della PR, Roberts P, et al. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open*. 2016;6:e011060.
22. Miller AR, Tucker C. Health information exchange, system size and information silos. *J Health Econ*. 2014;33:28–42.
23. Adler-Milstein J, DesRoches CM, Kralovec P, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. *Health Aff*. 2015;34:2174–2180.
24. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24:198–208.
25. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.
26. Silva Á, Cortez P, Santos MF, et al. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artif Intell Med*. 2006;36:223–234.

27. Gagne JJ, Glynn RJ, Avorn J, et al. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol*. 2011;64:749–759.
28. Tabak YP, Sun X, Nunez CM, et al. Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J Am Med Inform Assoc*. 2013;21:455–463.
29. Makar M, Ghassemi M, Cutler DM, et al. Short-term mortality prediction for elderly patients using Medicare claims data. *Int J Mach Learn Comput*. 2015;5:192–197.
30. Awad A, Bader-El-Den M, McNicholas J, et al. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inform*. 2017;108:185–195.
31. He D, Mathews SC, Kalloo AN, et al. Mining high-dimensional administrative claims data to predict early hospital readmissions. *J Am Med Inform Assoc*. 2014;21:272–279.
32. Kharrazi H, Chi W, Chang HY, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care*. 2017;55:789–796.
33. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. *Med Care*. 2018; 56:202–203.
34. Shadmi E, Balicer RD, Kinder K, et al. Assessing socioeconomic health care utilization inequity in Israel: impact of alternative approaches to morbidity adjustment. *BMC Public Health*. 2011;11:609.
35. Escobar GJ, Greene JD, Scheirer P, et al. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care*. 2008;46:232–239.
36. Escobar GJ, Ragins A, Scheirer P, et al. Nonelective rehospitalizations and postdischarge mortality: predictive models suitable for use in real time. *Med Care*. 2015;53:916–923.
37. Chen T, Guestrin C. Xgboost: A scalable tree boosting system: Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13–17, 2016. San Fransisco, CA; 2016:785–794.
38. Nguyen OK, Makam AN, Clark C, et al. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: model development and comparison. *J Hosp Med*. 2016; 11:473–480.
39. Tonkikh O, Shadmi E, Flaks-Manov N, et al. Functional status before and during acute hospitalization and readmission risk identification. *J Hosp Med*. 2016;11:636–641.
40. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the Electronic Cardiac Arrest Risk Triage (eCART) Score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg*. 2019;269:1059–1063.
41. Horne BD, Budge D, Masica AL, et al. Early inpatient calculation of laboratory-based 30-day readmission risk scores empowers clinical risk modification during index hospitalization. *Am Heart J*. 2017;185: 101–109.
42. Blumenthal D, Tavenner M. The “meaningful use” regulation for electronic health records. *N Engl J Med*. 2010;363:501–504.