HSR Health Services Research

DOI: 10.1111/1475-6773.12683
METHODS BRIEF

Improving Hospital Performance Rankings Using Discrete Patient Diagnoses for Risk Adjustment of Outcomes

Brendan DeCenso, Herbert C. Duber, Abraham D. Flaxman, Shane M. Murphy, and Michael Hanlon

Objective. To assess the changes in patient outcome prediction and hospital performance ranking when incorporating diagnoses as risk adjusters rather than comorbidity indices.

Data Sources. Healthcare Cost and Utilization Project State Inpatient Databases for New York State, 2005–2009.

Study Design. Conducted tree-based classification for mortality and readmission by incorporating discrete patient diagnoses as predictors, comparing with traditional comorbidity indices such as those used for Centers for Medicare and Medicaid Services (CMS) outcome models.

Principal Findings. Diagnosis codes as predictors increased predictive accuracy 5.6 percent (95% CI: 4.5–6.9 percent) relative to CMS condition categories for heart failure 30-day mortality. Most other outcomes exhibited statistically significant accuracy gains and facility ranking shifts. Sensitivity analysis showed improvements even when predictors were limited to only the diagnoses included in CMS models.

Conclusions. Discretizing patient severity information beyond the levels of traditional comorbidity indices improves patient outcome predictions and substantially shifts facility rankings.

Key Words. Risk adjustment, machine learning, Medicare

In an ongoing effort to incentivize quality health care delivery, the Centers for Medicare and Medicaid Services (CMS) has introduced two pay-for-performance initiatives monitoring hospital patient outcomes: (1) the Hospital Value-Based-Purchasing (HVBP) program, which will link 0.5 percent of FY17 Medicare DRG-based payments to 30-day mortality measures (Centers for Medicare and Medicaid Services, 2015), and (2) the Hospital Readmissions Reduction Program (HRRP), which links 3.0 percent of DRG-based

payments to readmission rates (Joynt and Jha 2013). For FY14, roughly one-quarter of eligible hospitals experienced a 0.2 percent or more reduction in payments under HVBP (Conway 2014), while approximately two-thirds saw penalties under HRRP (Boccuti and Casillas 2015). To control for varying average patient health ("risk") across facilities, CMS risk adjusts outcome measures using the condition categories forming Hierarchical Condition Categories (HCCs), a comorbidity index implemented in the early 2000s to predict expenditures and determine capitated payments for Medicare Advantage issuers (Pope et al. 2004). HCCs have been compared against baseline risk adjustment indices, such as pure age–sex models and the Elixhauser comorbidity index, and been found to be more representative of patient severity for major cardiovascular conditions (Li, Kim, and Doshi 2010; Krumholz et al. 2007).

To assess health facility performance, traditional risk adjustment of outcomes relies on hierarchical logistic regression with facility random effects. Defining predictors for such models involves grouping a large number of binary indicators, typically diagnosis codes in patient records, into a smaller set of medically relevant categories, such as HCCs (Krumholz et al. 2007). For instance, the presence of malignant neoplasms in respiratory or digestive organs registers as a single binary indicator representing cancers of those systems. Facility risk-adjusted rates are then calculated by inflating or deflating the mean population outcome by facility random effects, independent of variation explained by patient-level predictors.

Over the past two decades, data scientists have developed sophisticated nonparametric machine learning techniques as alternatives to logistic regression (James et al. 2013). Among these techniques are classification and regression trees (CART), which trace observations through single predictors at a time in a decision tree fashion. A main advantage of CART is that, unlike methods used to fit logistic regression, it does not estimate parameters simultaneously, computationally accommodating a much larger number of predictors. In terms of risk adjustment, this feature allows for diagnostic predictors to be kept discrete, avoiding information loss that could occur in grouping

---

Address correspondence to Brendan DeCenso, M.P.H., University of Pittsburgh School of Medicine, M240 Scaife Hall, 3550 Terrace St, Pittsburgh, PA 15261; e-mail: bmd70@pitt.edu. Herbert C. Duber, M.D., M.P.H., is with the Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, and also with the Division of Emergency Medicine, University of Washington, Seattle, WA. Abraham D. Flaxman, Ph.D., and Michael Hanlon, Ph.D., are with the Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA. Shane M. Murphy, Ph.D., is with the Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN.

them. Additionally, CART implicitly explores interactions between predictors, a flexibility that logistic modeling lacks (Touw et al. 2013). Past studies have examined nonparametric machine learning for risk adjustment and found improvements in accuracy over traditional methods (DiRusso et al. 2000; Eftekhar et al. 2005; Robinson 2008; Austin et al. 2012; Liu et al. 2015). Others have replicated CMS outcome models and compared to alternative predictor sets: Li, Kim, and Doshi (2010) found that HCCs outperform the Charlson and Elixhauser comorbidity indices; and Silber et al. (2010, 2016) found that including facility-level information, particularly patient volume, improved accuracy of the CMS outcome models. We uniquely combine aspects from the above studies by replicating CMS methodology, and comparing the predictive performance of CMS risk adjusters to that of discrete ICD-9-CM codes. We hypothesize that discretizing patient severity information will improve the accuracy of risk-adjusted outcomes to such an extent that hospital performance rankings will shift toward a truer order, implying that there exists a more appropriate allocation of performance-based payments.

METHODS

Data Source and Study Population

We identified cases of acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN) in New York State (NY) from 2006 to 2007 within the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID). We categorized the occurrences of each condition by applying CMS criteria to a patient's primary diagnosis (Grady et al. 2013a, b). We removed all the patients under the age of 66 to approximate a Medicare patient population with at least a one-year history. Other sample exclusions followed CMS methodology (Grady et al. 2013a, b), a listing of which can be found in Table 1. We were unable to precisely match the CMS pay-for-performance population in that we (1) excluded Medicare beneficiaries under age 65, and (2) included beneficiaries enrolled in Medicare Advantage, Medicare Hospice programs, or Veterans Health Administration Hospice programs in the 12 months prior to admission.

Outcomes

We examined nine outcomes: 30-day readmission, 30-day mortality, and in-facility mortality for each of AMI, HF, and PN. Our dataset consisted of NY

Table 1: Sample Exclusions, Demographic Information, and Crude Outcome Rates

	AMI			*HF*			*PN*		
	Mortality		*Readmission*	*Mortality*		*Readmission*	*Mortality*		*Readmission*
	In-facility	*30-day*	*30-day*	*In-facility*	*30-day*	*30-day*	*In-facility*	*30-day*	*30-day*
Admissions with condition as primary diagnosis at originating facility	75,904	75,904	75,904	141,842	141,842	141,842	132,882	132,882	132,882
Admissions—ages 66 and older	45,034	45,034	45,034	104,627	104,627	104,627	78,279	78,279	78,279
Index admission exclusions (among ages 66 and older)[†]									
Transferred without matching diagnosis at receiving facility	254	254	254	252	252	252	289	289	289
No visit linkage (patient history not available)	513	513	513	1,272	1,272	1,272	1,050	1,050	1,050
Discharged against physician advice	301	301	301	866	866	866	411	411	411
Age, sex, race[‡], or admission source missing	4,023	4,023	4,023	7,886	7,886	7,886	5,467	5,467	5,467
Residential zip code invalid or n/a[‡]	328	328	328	579	579	579	480	480	480
Missing income data[‡]	3,135	3,135	3,135	3,455	3,455	3,455	3,313	3,313	3,313
Admitted and discharged on same day (likely not clinically significant)	1,237	1,237	1,237	732	732	732	685	685	685
Outcome-specific index admission exclusions[†]									
Deaths assigned to original facility in case of transfer	7,972	7,972		2,060	2,060		717	717	

Continued

Table 1 *Continued*

	AMI			*HF*			*PN*		
	Mortality		*Readmission*	*Mortality*		*Readmission*	*Mortality*		*Readmission*
	In-facility	*30-day*	*30-day*	*In-facility*	*30-day*	*30-day*	*In-facility*	*30-day*	*30-day*
Readmissions within 30 days not considered index admissions			5,251			22,592			11,455
In-facility deaths not considered index admissions			4,397			5,346			6,218
Readmissions assigned to final facility in case of transfer			6,654			3,545			1,597
Readmissions occurring within 1 day of discharge considered transfers			1,914			460			351
One admission randomly selected for patients w>1 admission during calendar year	2,112	2,112	699	21,259	21,259	9,646	5,991	5,991	2,783
Indeterminate 30-day mortality		8,406			15,629			17,347	
Total admissions	29,497	21,125	26,293	70,786	55,234	60,855	63,933	46,673	54,238
Random forest train set (2006)	16,387	12,000	14,417	38,213	30,275	32,865	34,993	26,203	29,647
Model set (2007)	13,110	9,125	11,876	32,573	24,959	27,990	28,940	20,470	24,591
Population characteristics (based on total admissions)									
Facility count	195	194	194	208	208	209	210	209	211
Outcome rate (%)	10.4	16.2	21.6	5.0	8.6	23.4	7.5	12.1	18.1
Mean age	80.2	80.5	79.6	80.7	80.7	80.8	80.7	80.7	80.6
Female (%)	54.0	54.8	52.6	56.7	57.0	57.1	54.7	54.4	55.3

†Values represent the number of observations excluded for each rule (note that the sum of these values subtracted from admissions for ages 66+ will be less than the final sample, as some observations failed to meet multiple criteria).

‡Socioeconomic characteristics factored into a version of our analysis not included in this paper, and as such, we removed individuals missing socioeconomic information.

inpatient records, from which we captured 30-day readmission provided the readmission occurred in New York. We did not count as readmissions those admissions with diagnoses or procedures defined by CMS as planned visits (Grady et al. 2013b). Transfers were treated as single observations by assigning deaths to the originating hospital and readmissions to the final hospital. Because our data lacked vital registration information, we could not directly assign 30-day mortality. To account for this limitation, we referenced 2006–2009 NY admissions and deaths in the inpatient setting to assign 30-day mortality for 76.4 percent of 2006 admissions and 73.1 percent of 2007 admissions in our analysis. We then dropped from the 30-day mortality sample admissions with indeterminate status—those corresponding to out-of-facility deaths, or patients not admitted to a NY facility between 30 days following their index admission and year-end 2009. For the sake of having a complete mortality outcome, we also examined in-facility mortality. It should be noted that in-facility mortality produces biased performance rankings in favor of facilities with low length-of-stay (Rosenthal et al. 2000).

Predictors

We compiled seven predictor sets: (1) age and sex alone ("Age–sex"), (2) Elixhauser comorbidities ("Elixhauser"; Elixhauser et al. 1998), (3) CMS condition categories used for HCCs ("CMS"), (4) three-digit ICD-9-CM diagnosis codes ("3-digit ICD expanded"), (5) a restricted set of three-digit ICD-9-CM diagnosis codes ("3-digit ICD restricted"), (6) five-digit ICD-9-CM diagnosis codes ("5-digit ICD expanded"), and (7) a restricted set of five-digit ICD-9-CM diagnosis codes ("5-digit ICD restricted"). The "restricted" ICD predictor sets consisted of only diagnosis codes that translate into CMS condition categories used for a given outcome—that is to say, the only difference between the CMS predictor set and the ICD "restricted" predictor sets was the extent to which patient diagnostic information was discretized. Per 2008 CMS model revisions (Bhat et al. 2008), we did not impose hierarchies on condition categories to form HCCs. We combined related condition categories in line with CMS methodology (e.g., schizophrenia and major depressive disorder) and reduced categories to only those used by CMS models. For the ICD predictor sets, binary indicators represented the presence of specific three-digit (e.g., 276.XX) or five-digit (e.g., 276.51) diagnosis codes. We populated the Elixhauser, CMS, and ICD predictor sets with one-year inpatient histories, and from index admissions included only diagnoses present-on-admission, as established in the literature (Iezzoni 2007). Notably, CMS excludes diagnoses

for condition categories indicating complications of the index admission even if present-on-admission (Grady et al. 2013a, b)—we removed all such diagnoses from both the CMS and the ICD predictor sets. We additionally removed five diagnoses from the ICD "expanded" predictor sets (995.91, 995.92, 348.1, 780.01, and 783.7) that could reasonably be facility induced, do not translate to HCCs, and were found to be highly predictive of outcomes. Finally, we removed from the ICD "expanded" predictor sets all ICD-9-CM V codes that do not translate to CMS condition categories, as V codes are less often tied to payment and therefore vary in usage across facilities. Appendix SA3 provides further detail on ICD-9-CM codes excluded from the ICD "expanded" predictor sets.

Tree-Based Classification

For outcome predictions, we utilized random forests, a classification technique that incorporates uncertainty in two ways: (1) by creating multiple decision trees with bootstrapped datasets drawn from a training set of observations, and (2) by randomly selecting only a subset of all predictors at each node in a tree, and subgrouping observations by the predictor among the selected subset that yields the best split of the data (Breiman 2001). For classification, the "best split of the data" at a node minimizes heterogeneity in outcomes for immediately subsequent groupings. Once all splits have been made and decision trees compiled, test set observations are run through the splitting rules of each tree, with the expectation of a test observation having a particular outcome equaling the fraction of trees assigning the observation to that outcome.

For each outcome and predictor set, we first trained random forests on 2006 admissions. Each of these training runs provided two outputs of interest: (1) predictions ("risk scores") for 2007 admissions, to be used in the logistic modeling described below, and (2) an "importance" score for each predictor, representing the extent to which accuracy decreased if the predictor was left out of classification. We used the randomForest package in *R* version 3.3.1 for classification (R Core Team 2016; Liaw and Wiener 2002).

Model Comparisons

To produce model fit statistics—the c-statistic and Hosmer–Lemeshow chi-square value (Hosmer and Lemeshow 1980)— we ran 10-fold cross validation on 2007 admissions with logistic regression fit by maximum likelihood

estimation without facility random effects. The risk scores outputted by random forests for 2007 admissions served as the only right-hand-side variable.

We then used hospital-clustered bootstrapping to generate 500 datasets consisting of 2007 admissions for each outcome and calculated the c-statistic as described above for each predictor set on each bootstrapped dataset. Also for each predictor set on each bootstrapped dataset, we ran mixed effects logistic regression fit by maximum likelihood with facility random effects. The risk scores outputted by random forests for each predictor set served as fixed effects in these models. We then calculated facility risk-adjusted rates dividing predicted outcomes (facility random effect included) by expected outcomes (facility random effect excluded) and multiplying by the mean outcome in the full sample. Within each bootstrapped dataset, we ranked facilities according to their risk-adjusted rates and calculated the change in ranking between the CMS predictor set and other predictor sets. We then calculated across all bootstraps each facility's mean risk-adjusted rate, the average change in ranking and average relative percent change in c-statistic between the CMS predictor set and other predictor sets, as well as the 2.5th percentile and 97.5th percentile change in ranking and relative percent change in c-statistic between the CMS predictor set and the other predictor sets. We further described reclassification under the five-digit ICD "restricted" and "expanded" predictor sets relative to the CMS predictor set using facility ranking scatterplots (based on facility mean risk-adjusted rates) and quintiles of risk-adjusted rates. We conducted all data processing, logistic modeling, and analysis in *Stata SE* version 14.2 (StataCorp 2015).

RESULTS

Table 1 presents demographic information, sample counts, and crude rates by outcome. Table 2 displays model fit statistics for each outcome and predictor set, as well as facility ranking changes and c-statistic changes relative to the CMS predictor set. The five-digit ICD predictors outperformed CMS predictors for 30-day mortality and 30-day readmission outcomes, with anywhere from a 1.5–3.5 and 1.7–5.6 relative percent increase in the c-statistic over CMS predictors for the "restricted" and "expanded" sets, respectively. The five-digit ICD predictors also outperformed corresponding three-digit ICD predictors.

Appendix SA2 lists the 200 most important predictors for the five-digit ICD "restricted" and "expanded" predictor sets. For the most part, conditions

Table 2: Model Statistics

Outcome	Predictor Set	Mean Facility Spots Changed in Rankings Relative to CMS (95% CI)*†			Percent Change in C-statistic Relative to CMS (95% CI)‡			C-statistic			H-L Chi-Square		
		AMI	HF	PN	AMI	HF	PN	AMI	HF	PN	AMI	HF	PN
In-facility mortality	Age–sex	9.3 (2.0–17.5)	6.0 (0.0–13.1)	6.5 (0.6–13.4)	−8.6 (−10.9 to −6.2)	−9.0 (−10.9 to −7.1)	−10.2 (−11.9 to −8.5)	0.614	0.596	0.597	11.7	12.5	10.4
	Elixhauser	8.4 (1.2–16.4)	6.9 (0.4–14.3)	7.2 (0.7–14.3)	−1.8 (−4.3 to 0.8)	−2.1 (−4.1 to 0.2)	−1.9 (−3.5 to −0.2)	0.659	0.642	0.653	14.6	20.6	22.3
	CMS	–	–	–	–	–	–	0.671	0.655	0.665	15.1	15.0	22.3
	3-digit ICD (restricted)	2.8 (−1.1 to 7.9)	4.7 (−0.6 to 11)	3.4 (−0.1 to 8.7)	0.8 (0.0–1.9)	−0.4 (−1.6 to 0.9)	0.5 (−0.7 to 1.5)	0.676	0.653	0.668	18.8	13.6	19.1
	3-digit ICD (expanded)	8.1 (1.3–15.8)	9.3 (1.8–17.8)	8.3 (1.4–16.0)	3.1 (1.4–4.9)	3.2 (1.6–5.0)	1.5 (0.2–2.9)	0.692	0.676	0.675	17.6	21.4	27.6
	5-digit ICD (restricted)	5.1 (−0.3 to 11.6)	6.4 (0.2–13.5)	5.2 (−0.1 to 11.3)	1.8 (0.6–3.1)	1.3 (−0.2 to 3.0)	1.7 (0.5–2.8)	0.683	0.664	0.676	15.6	12.4	13.7
	5-digit ICD (expanded)	9.6 (2.0–17.9)	10.2 (2.4–19.0)	8.2 (1.4–15.8)	4.9 (3.3–6.7)	6.2 (4.4–7.9)	4.4 (3.2–5.7)	0.703	0.695	0.694	18.2	17.7	18.9
30-day mortality	Age–sex	9.9 (2.2–18.4)	7.4 (1.1–14.7)	9.5 (2.4–17.2)	−8.2 (−10.1 to −6.2)	−8.1 (−9.7 to −6.7)	−11.8 (−13.6 to −10.1)	0.613	0.610	0.592	10.5	12.9	18.0
	Elixhauser	9.7 (1.9–18.1)	8.4 (1.6–16.0)	8.5 (1.7–16)	−2.1 (−4.5 to 0.4)	−2.0 (−3.7 to −0.6)	−2.8 (−4.4 to −1.2)	0.653	0.650	0.653	12.0	16.1	16.9
	CMS	–	–	–	–	–	–	0.668	0.663	0.672	11.6	14.3	12.1
	3-digit ICD (restricted)	3.1 (−0.9 to 8.4)	4.8 (−0.3 to 10.9)	5.0 (−0.1 to 11.1)	0.5 (−0.4 to 1.5)	0.1 (−1.1 to 1.1)	0.6 (−0.1 to 1.4)	0.671	0.663	0.676	11.8	13.4	13.0
	3-digit ICD (expanded)	8.6 (1.4–16.7)	9.2 (1.8–17.5)	9.8 (2.5–17.8)	3.1 (1.2–5.0)	2.6 (1.2–4.0)	2.0 (0.8–3.3)	0.688	0.681	0.685	13.6	19.8	19.8
	5-digit ICD (restricted)	6.1 (0.2–13.2)	6.2 (0.4–13.0)	7.0 (0.9–14.0)	1.5 (0.3–2.8)	1.5 (0.4–2.8)	1.8 (0.7–2.7)	0.677	0.674	0.683	12.3	10.5	8.4

Continued

Table 2 *Continued*

Outcome	Predictor Set	Mean Facility Spots Changed in Rankings Relative to CMS (95% CI)*†			Percent Change in C-statistic Relative to CMS (95% CI)‡			C-statistic			H-L Chi-Square		
		AMI	HF	PN	AMI	HF	PN	AMI	HF	PN	AMI	HF	PN
	5-digit ICD (expanded)	9.6 (2.1–18.1)	9.5 (2.0–17.7)	11.3 (3.3–20.0)	4.2 (2.5–6.0)	5.6 (4.5–6.9)	4.0 (2.7–5.4)	0.695	0.700	0.699	13.8	15.4	13.1
30-day readmission	Age–sex	6.8 (0.1–14.6)	11.2 (3.0–20.3)	11.6 (3.7–20.2)	−15.2 (−17.5 to −12.7)	−22.4 (−25.0 to −20.8)	−20.0 (−21.6 to −18.4)	0.523	0.507	0.515	9.6	6.5	9.0
	Elixhauser	6.2 (−0.1 to 13.4)	10.6 (2.3–19.4)	9.3 (2.1–17.2)	−7.8 (−10.0 to −5.8)	−8.5 (−9.8 to −7.1)	−8.4 (−9.8 to −7.1)	0.568	0.596	0.590	9.1	8.6	15.0
	CMS	–	–	–	–	–	–	0.616	0.652	0.644	9.1	15.7	11.9
	3-digit ICD (restricted)	2.3 (−1.5 to 7.0)	5.5 (−0.4 to 12.2)	5.7 (0.2–11.9)	0.6 (−0.1 to 1.2)	0.6 (0.0 to 1.2)	2.4 (1.6 to 3.2)	0.619	0.656	0.660	7.4	10.5	9.3
	3-digit ICD (expanded)	3.6 (−1.0 to 9.4)	5.2 (−0.5 to 11.9)	6.2 (0.5–12.8)	0.4 (−0.8 to 1.6)	1.9 (1.1–2.6)	2.3 (1.3–3.3)	0.619	0.664	0.659	8.3	8.5	8.3
	5-digit ICD (restricted)	3.7 (−1.2 to 9.4)	6.9 (−0.1 to 15.1)	7.6 (1.0–15.0)	1.6 (0.5–2.6)	2.2 (1.6–2.8)	3.5 (2.4–4.5)	0.625	0.666	0.667	8.7	7.1	9.1
	5-digit ICD (expanded)	4.5 (−0.8 to 10.9)	7.3 (0.3–15.2)	9.0 (2.1–16.6)	1.7 (0.5–2.7)	3.2 (2.4–3.9)	3.4 (2.3–4.5)	0.626	0.672	0.666	9.1	9.9	9.2

*In the first six columns, bootstrapped statistics are presented. For the first three columns, we first calculated the mean, 2.5th percentile, and 97.5th percentile difference in rankings from CMS rankings for each facility across all bootstraps. We then normalized those values such that for each facility, the mean difference was set to be positive. We then averaged the mean, 2.5th percentile, and 97.5th percentile values across all facilities. A negative value for the 2.5th percentile indicates that the 2.5th percentile of change in ranking for the average facility was in the opposite direction as the mean change in ranking.

†Number of facilities: AMI in-facility mort—195; AMI 30-day mort—194; AMI 30-day read—194; HF in-facility mort—208; HF 30-day mort—208; HF 30-day read—209; PN in-facility mort—210; PN 30-day mort—209; PN 30-day read—211.

‡Percent change = [(predictor set c-statistic − CMS c-statistic)/CMS c-statistic] × 100.

that served as strong predictors across all outcomes—for instance, chronic kidney disease, malignant neoplasms, and dementia—appeared in both "restricted" and "expanded" predictor sets, indicating they are conditions included in CMS outcome models. Table 3 presents the most important five-digit ICD "expanded" predictors not included in CMS models for each outcome.

Scatterplots of five-digit ICD facility rankings against CMS rankings are shown in Figure 1. When facility risk-adjusted rates were grouped into quintiles, 26.8, 21.2, and 30.1 percent of facilities shifted from one quintile to another under the five-digit ICD "expanded" predictor set relative to the CMS predictor set for AMI, HF, and PN 30-day mortality, respectively. Similarly, 12.4, 19.1, and 22.7 percent of facilities shifted from one quintile to another for AMI, HF, and PN 30-day readmission, respectively. For the five-digit ICD "restricted" predictor sets, these values were 17.0, 15.4, and 21.1 percent for AMI, HF, and PN 30-day mortality, respectively; and 12.4, 18.2, and 18.0 percent for AMI, HF, and PN 30-day readmission, respectively.

Discussion

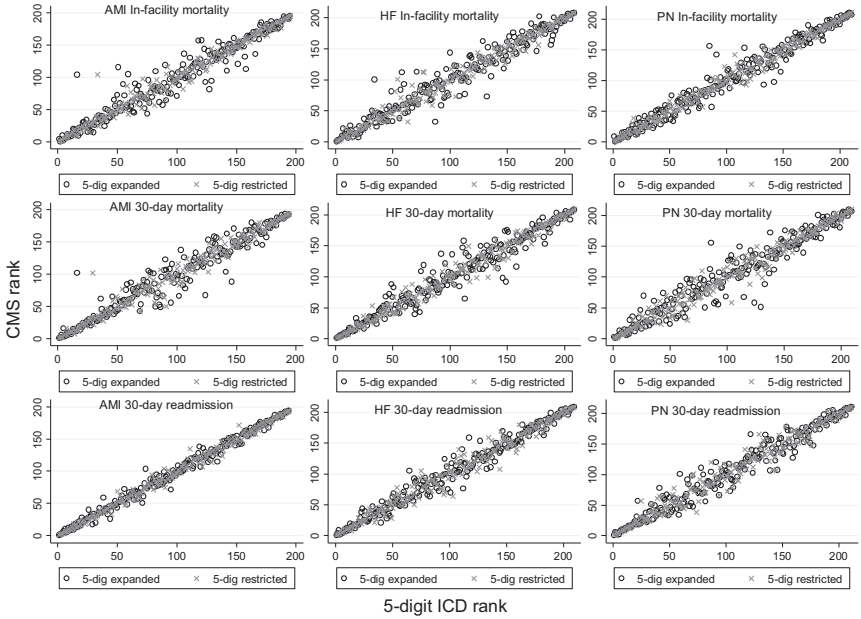
We used tree-based classification to predict mortality and readmission for three different conditions, finding that individual ICD-9-CM diagnoses yielded more accurate predictions for all outcomes relative to CMS condition categories. This relationship held even for our five-digit ICD "restricted" predictors, where the only difference from CMS condition categories was that individual diagnoses were broken out rather than grouped together. This finding that discretization of patient severity information (beyond traditional levels) improves outcome prediction is further evidenced by the five-digit ICD predictors outperforming three-digit ICD predictors.

We additionally found substantial facility ranking shifts when comparing the five-digit ICD "expanded" predictors and CMS predictors across all outcomes, with the average facility moving in the rankings by roughly 5 percent for most outcomes, and up to 30 percent of facilities shifting into a different quintile of performance. Notably, ranking shifts observed when moving from CMS predictors to five-digit ICD "expanded" predictors in some cases exceeded the shifts observed when moving from the purely demographic age–sex predictors to CMS predictors. When combined with the increased accuracy of ICD predictors, these results point to the strengths of a data-driven approach—ICD predictors outperformed CMS predictors and sizably

Table 3: Most Important Five-Digit ICD-9-CM Diagnoses Not Included in CMS Outcome Models (Importance Ranking among All Five-Digit ICD-9-CM Diagnoses in Parentheses)

	AMI		*HF*		*PN*	
In-facility mortality	286.9	Coagulat defect NEC/NOS (3)	E878.8	Abn react-surg proc NEC (5)	787.2	Dysphagia (3)
	790.9	Abnrml coagultion prfile (11)	787.2	Dysphagia (7)	E878.8	Abn react-surg proc NEC (4)
	288.8	Wbc disease NEC (13)	518.0	Pulmonary collapse (8)	707.14	Ulcer of heel & midfoot (5)
	518.0	Pulmonary collapse (15)	E879.1	Abn react-renal dialysis (13)	424.1	Aortic valve disorder (7)
	285.9	Anemia NOS (18)	287.5	Thrombocytopenia NOS (14)	707.15	Ulcer other part of foot (10)
	421.0	Ac/subac bact endocard (20)	E879.8	Abn react-procedure NEC (15)	707.03	Pressure ulcer, low back (14)
	789.5	Ascites (21)	284.8	Oth spec aplastic anemias (16)	397.0	Tricuspid valve disease (18)
	780.4	Convulsions NEC (23)	682.6	Cellulitis of leg (18)	560.1	Paralytic ileus (19)
	560.1	Paralytic ileus (26)	786.09	Respiratory abnorm NEC (19)	287.5	Thrombocytopenia NOS (24)
	287.5	Thrombocytopenia NOS (27)	008.45	Int inf clstrdium dfcile (22)	593.9	Renal & ureteral dis NOS (26)
30-day mortality	286.9	Coagulat defect NEC/NOS (7)	787.2	Dysphagia (9)	199.1	Malignant neoplasm NOS (9)
	285.2	Anemia in neoplastic dis (16)	E878.8	Abn react-surg proc NEC (11)	287.5	Thrombocytopenia NOS (10)
	288.8	Wbc disease NEC (18)	286.9	Coagulat defect NEC/NOS (12)	707	Pressure ulcer, low back (12)
	421.0	Ac/subac bact endocard (19)	293	Delirium d/t other cond (13)	348.3	Metabolic encephalopathy (13)
	682.6	Cellulitis of leg (22)	599.7	Hematuria (14)	783.7	Failure to thrive-adult (15)
	790.9	Abnrml coagultion prfile (26)	276.2	Acidosis (15)	787.2	Dysphagia (18)
	787.9	Diarrhea (32)	515	Postinflam pulm fibrosis (16)	273.8	Dis plas protein met NEC (25)
	285.9	Anemia NOS (34)	788.9	Oth symp urinary system (17)	424.1	Aortic valve disorder (26)
	560.1	Paralytic ileus (38)	285.2	Anemia in chr kidney dis (19)	707.1	Pressure ulcer, heel (28)
	287.5	Thrombocytopenia NOS (41)	518.0	Pulmonary collapse (20)	348.3	Encephalopathy NOS (31)
30-day readmission	518.81	Acute respiratory failure (7)	599.0	Urin tract infection NOS (4)	787.2	Dysphagia (8)
	787.91	Diarrhea (14)	788.20	Retention urine NOS (12)	786.59	Chest pain NEC (11)
	518.89	Other lung disease NEC (15)	682.7	Cellulitis of foot (20)	783.7	Failure to thrive-adult (17)
	300.00	Anxiety state NOS (17)	284.8	Oth spec aplastic anemias (23)	578.9	Gastrointest hemorr NOS (18)
	564.00	Constipation NOS (18)	780.39	Convulsions NEC (26)	008.45	Int inf clstrdium dfcile (22)
	E879.6	Abn react-urinary cath (20)	286.9	Coagulat defect NEC/NOS (27)	786.50	Chest pain NOS (23)
	562.10	Dvrtclo colon w/o hmrhg (24)	511.9	Pleural effusion NOS (30)	560.39	Impaction intestine NEC (25)
	535.50	Gstr/ddnts NOS w/o hmrhg (26)	600.01	BPH w urinary obs/LUTS (31)	251.8	Pancreatic disorder NEC (27)
	599.0	Urin tract infection NOS (27)	799.3	Debility NOS (41)	278.00	Obesity NOS (38)
	2,962.0	Depress psychosis-unspec (34)	780.99	Other general symptoms (44)	276.5	Volume depletion disorder (41)

Figure 1: Performance Rankings, Five-Digit ICD Relative to CMS



impacted performance rankings. The introduction of ICD-10-CM in the United States, with its four-fold increase in diagnosis codes, presents an opportunity to further leverage these gains in predictive accuracy.

As shown in Table 3, results from our five-digit ICD "expanded" models yielded diagnoses that were highly informative in classification models but might not be traditionally considered for risk adjustment. These diagnoses included sequelae of serious conditions (e.g., anemia in neoplastic/chronic kidney disease), signs of underlying debility and/or weakened immune response (e.g., cellulitis, abnormal coagulation, obesity, dysphagia, diarrhea, a history of being coded with "failure to thrive"), and disorders of systems tangentially related to the condition of interest (e.g., aortic valve disorder for PN outcomes, gastrointestinal inflammation/bleeding for all outcomes). This ability to identify predictors, which would not at first glance seem to be clinically relevant to the outcome of interest, is another benefit of a nonparametric machine learning approach.

We note several limitations in our analysis. First, despite our efforts to exclude potentially facility-induced codes from our ICD "expanded" models, there are likely still some of these codes among the strongest predictors. For

the sake of transparency, we have included a list of the top-200 strongest predictors in Appendix SA2. An additional limitation is that we utilized classification trees for all predictor sets to facilitate comparisons across their outputs and parse out the incremental effect of discretizing patient severity information. Multiple studies have shown little advantage or even a disadvantage for implementing CART over logistic regression when run on identical predictor sets (Colombet et al. 2000; Dreiseitl and Ohno-Machado 2002; Terrin et al. 2003; Austin 2007). Hierarchical logistic regression and/or more fine-tuned tree-based classification may be better modeling tools for the objective of this analysis and merit further exploration. Additional limitations in our analysis include (1) our 30-day mortality assignment was incomplete (we acknowledge that patients dropped due to unknown 30-day mortality were likely those for whom outcomes would be most difficult to predict; however, all models compared in this analysis were subject to this same bias), (2) we populated patient histories using only encounters from the inpatient setting, and (3) our sample did not precisely capture the population used for CMS hospital performance measures.

Despite the above limitations, we believe that incorporating individual patient diagnoses as severity predictors demonstrates strong potential to improve outcome prediction and hospital performance rankings. The results presented here have relevance for U.S. performance-based financing programs, such as HVBP and HRRP, in that even small ranking shifts under those programs can be enough to impact facility payments. To clarify these findings, future analyses could quantify the degree to which using different risk adjustment models influences hospital performance-based payments, as well as assess the potential to improve patient outcome prediction with the further discretization of diagnostic information under ICD-10-CM.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was funded in part by the Bill & Melinda Gates Foundation through research conducted at the Institute for Health Metrics and Evaluation within the University of Washington Department of Global Health. The funder took no part in conceptual design, analysis, or manuscript preparation, nor did the funder participate in the decision to submit for publication.

Disclosures: None.

Disclaimer: None.

References

Austin, P. C. 2007. "A Comparison of Regression Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines for Predicting AMI Mortality." *Statistics in Medicine* 26 (15): 2937–57.

Austin, P. C., D. S. Lee, E. W. Steyerberg, and J. V. Tu. 2012. "Regression Trees for Predicting Mortality in Patients with Cardiovascular Disease: What Improvement Is Achieved by Using Ensemble-Based Methods?" *Biometrical Journal.* 54 (5): 657–73.

Bhat, K. R., E. E. Drye, H. M. Krumholz, S.-L. T. Normand, G. C. Schreiner, Y. Wang, and Y. Wang. 2008. "Acute Myocardial Infarction, Heart Failure, and Pneumonia Mortality Measures Maintenance." Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (Yale-CORE) [accessed on May 2, 2013]. Available at http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier3&cid=1163010421830

Boccuti, C., and G. Casillas. 2015. "Aiming for Fewer Hospital U-Turns: The Medicare Hospital Readmission Reduction Program" [accessed on June 25, 2015]. Available at http://kff.org/medicare/issue-brief/aiming-for-fewer-hospital-u-turns-the-medicare-hospital-readmission-reduction-program/

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Centers for Medicare and Medicaid Services. 2015. "Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System Policy Changes and Fiscal Year 2016 Rates; Revisions of Quality Reporting Requirements for Specific Providers, Including Changes Related to the Electronic Health Record Incentive Program; Extensions of the Medicare-Dependent, Small Rural Hospital Program and the Low-Volume Payment Adjustment for Hospitals." Federal Register Vol 80 No 158. 42 CFR Part 412, pp. 49325–886.

Colombet, I., A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M. C. Jaulent. 2000. "Models to Predict Cardiovascular Risk: Comparison of CART, Multilayer Perceptron and Logistic Regression." *Proceedings of the AMIA Symposium*, 156–60.

Conway, P. 2014. "CMS Releases Latest Value-Based Purchasing Program Scorecard. The CMS Blog" [accessed on June 2, 2015]. Available at http://blog.cms.gov/2013/11/14/cms-releases-latest-value-based-purchasing-program-scorecard/

DiRusso, S. M., T. Sullivan, C. Holly, S. N. Cuff, and J. Savino. 2000. "An Artificial Neural Network as a Model for Prediction of Survival in Trauma Patients: Validation for a Regional Trauma Area". *The Journal of Trauma–Injury Infection and Critical Care* 49 (2): 212–20.

Dreiseitl, S., and L. Ohno-Machado. 2002. "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." *Journal of Biomedical Informatics* 35 (5–6): 352–9.

Eftekhar, B., K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi. 2005. "Comparison of Artificial Neural Network and Logistic Regression Models for

Prediction of Mortality in Head Trauma Based on Initial Clinical Data." *BMC Medical Informatics and Decision Making* 5 (1): 3.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36 (1): 8–27.

Grady, J. N., Z. Lin, Y. Wang, M. Keenan, K. R. Bhat, H. M. Krumholz, and S. M. Bernheim. 2013a. "2013 Condition-Based Measure Updates and Specifications: Acute Myocardial Infarction, Heart Failure, and Pneumonia 30-Day Risk-Standardized Mortality Measures" [accessed on August 19, 2013]. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Available at https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1163010421830.

Grady, J. N., Z. Lin, C. Wang, M. Keenan, C. Nwosu, K. R. Bhat, L. I. Horwitz, E. E. Drye, H. M. Krumholz, and S. M. Bernheim. 2013b. "2013 Measures Updates and Specifications Report: Hospital-Level 30-Day Risk-Standardized Readmission Measures for Acute Myocardial Infarction, Heart Failure, and Pneumonia" [accessed on August 19, 2013]. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html

Hosmer, D. W., and S. Lemeshow. 1980. "Goodness of Fit Tests for the Multiple Logistic Regression Model." *Communications in Statistics – Theory and Methods* 9 (10): 1043–69.

Iezzoni, L. I. 2007. "Finally Present on Admission but Needs Attention." *Medical Care* 45 (4): 280–2.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Joynt, K. E., and A. K. Jha. 2013. "A Path Forward on Medicare Readmissions." *New England Journal of Medicine* 368 (13): 1175–7.

Krumholz, H. M., S. L. T. Norman, D. H. Galusha, J. A. Mattera, A. S. Rich, Y. Wang, and Y. Wang. 2007. "Risk-Adjustment Models for AMI and HF: 30-Day Mortality" [accessed on December 8, 2011]. Available at http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1163010421830

Li, P., M. M. Kim, and J. A. Doshi. 2010. "Comparison of the Performance of the CMS Hierarchical Condition Category (CMS-HCC) Risk Adjuster with the Charlson and Elixhauser Comorbidity Measures in Predicting Mortality." *BMC Health Services Research* 10 (1): 245.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Liu, Y., M. Traskin, S. A. Lorch, E. I. George, and D. Small. 2015. "Ensemble of Trees Approaches to Risk Adjustment for Evaluating a Hospital's Performance." *Health Care Management Science* 18 (1): 58–66.

Pope, G. C., J. Kautter, R. P. Ellis, A. S. Ash, J. Z. Ayanian, L. I. Iezzoni, M. I. Ingber, J. M. Levy, and J. Robst. 2004. "Risk Adjustment of Medicare Capitation

Payments Using the CMS-HCC Model." *Health Care Financing Review* 25 (4): 119–41.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robinson, J. W. 2008. "Regression Tree Boosting to Adjust Health Care Cost Predictions for Diagnostic Mix." *Health Services Research* 43 (2): 755–72.

Rosenthal, G. E., D. W. Baker, D. G. Norris, L. E. Way, D. L. Harper, and R. J. Snow. 2000. "Relationships between In-Hospital and 30-Day Standardized Hospital Mortality: Implications for Profiling Hospitals." *Health Services Research* 34 (7): 1449–68.

Silber, J. H., P. R. Rosenbaum, T. J. Brachet, R. N. Ross, L. J. Bressler, O. Even-Shoshan, S. A. Lorch, and K. G. Volpp. 2010. "The Hospital Compare Mortality Model and the Volume–Outcome Relationship." *Health Services Research* 45 (5 Pt 1): 1148–67.

Silber, J. H., V. A. Satopää, N. Mukherjee, V. Rockova, W. Wang, A. S. Hill, O. Even-Shoshan, P. R. Rosenbaum, and E. I. George. 2016. "Improving Medicare's Hospital Compare Mortality Model." *Health Services Research* 51 (3 Pt 2): 1229–47.

StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

Terrin, N., C. H. Schmid, J. L. Griffith, R. B. D'Agostino, and H. P. Selker. 2003. "External Validity of Predictive Models: A Comparison of Logistic Regression, Classification Trees, and Neural Networks." *Journal of Clinical Epidemiology* 56 (8): 721–9.

Touw, W. G., J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum. 2013. "Data Mining in the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle?" *Briefings in Bioinformatics* 14 (3): 315–26.

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Author Matrix.
Appendix SA2: Most Important Predictors in Five-Digit ICD Models.
Appendix SA3: Code Exclusions for ICD "Expanded" Sets.