Are Patient Risk Factors Consistent Across Data Sources: A Comparison of EMR, Billing, and Clinician-Abstracted Data

Presenting Author: Theodora Wingert, MD
Co-Authors: Ira S. Hofer, MD, Tristan Grogan, MS, Melissa D. McCabe, MD, Eilon Gabel, MD, Richard Shemin, MD, Aman Mahajan, MD, PhD, Maxime Cannesson, MD, PhD

Background: Widespread implementation of electronic medical records (EMR) and improved statistical techniques have created an explosion in the amount of health data available to researchers. Although these advances open new avenues for researchers, minimal endeavors have been made to establish quality and assess correlation between the various data sources. We aim to evaluate the consistency of data across administrative, research, and clinical data sources and hypothesize that data will be well correlated.

Methods: After IRB approval, patients who underwent cardiac surgery between April 1, 2013 and March 26, 2014 were identified from our institutional submission to the Society for Thoracic Surgery (STS) registry. Patient data from the "Risk Factor" and "Operative Description" sections of the STS database were mapped to one of 15 comorbidities (Table 1). EMR documentation of these comorbidities was obtained from the "Past Medical History" and "Problem List." Lastly, ICD-9 billing data were obtained by mapping codes to the 15 comorbidities. STS was considered the gold standard for statistical analysis as all data was clinician-abstracted and it was presumed to be the highest quality, sensitivities and specificities were calculated accordingly. Fleiss kappa coefficients were calculated to assess concordance between the data sets with adjustment for prevalence. In order to examine the impact of misclassification on risk stratification models, a linear regression analysis of a risk-adjusted length of stay (LOS) for patients using each of the three data sets was performed.

Results: The STS database contained 431 patients, 388 were successfully matched to corresponding records in the EMR. The prevalence of comorbidities varied across data sources, with a ten-fold difference in valvular disease and smoking history. Sensitivity ranged from a low of 7% for substance abuse to a high of >90% for diabetes and coronary artery disease, averaging 59% for both EMR and ICD-9 (Table 1). Specificity measures were higher, averaging 82% for EMR and 79% for ICD-9 (Table 1). Kappa coefficients ranged between -0.19 and 0.74, and were higher for EMR compared to STS than ICD-9 compared to STS (0.39 vs. 0.23). In our regression model using STS data only, history of arrhythmia and hyperlipidemia were found to be significantly associated with LOS. However in the model using EMR data, history of arrhythmia, cerebrovascular disease, and congestive heart failure (CHF) were associated with LOS, and using ICD-9 data, history of arrhythmia, CHF, substance abuse, diabetes, end-stage renal disease, hypertension, and chronic obstructive pulmonary disease were associated with LOS.

Discussion: Comparison of comorbidities between STS, EMR, and ICD-9 data demonstrated significant variability in prevalence and at best, fair agreement between the data sets. In contrast to previous reports of the reliability of ICD-9 data, we found data validity varied with the comorbidity evaluated. Similarly, comorbidities predictive of LOS varied based on the data set utilized for modeling. Although the relatively small sample size limits this retrospective study, we believe continued efforts to validate data accuracy are vitally important to the future of research studies using data obtained from administrative and billing sources.

Table 1.

	Prevalence (%)			EMR (%)		ICD9 (%)		Fleiss kappa coefficient	
	STS	EMR	ICD9	Sensitivity	Specificity	Sensitivity	Specificity	EMR	ICD9
Diabetes Mellitus	29.1	37.1	65.5	93.8	86.2	64.1	74.5	0.738	0.147
Valvular Disease	51.3	54.9	5.4	79.3	47.1	20.7	95.8	0.721	-0.193
Coronary Artery Disease	45.9	58.0	27.1	96.2	56.2	76.4	50.2	0.644	0.271
Obstructive Sleep Apnea	7.5	10.1	11.3	75.9	95.3	79.3	94.2	0.613	0.592
End Stage Renal Disease	5.2	7.7	8.8	75.0	95.9	65.0	94.3	0.572	0.443
Hypertension	51.0	59.3	72.2	81.3	63.7	84.8	41.1	0.448	0.226
Cerebrovascular Disease	8.5	13.7	18.0	63.6	91.0	67.6	88.0	0.425	0.34
Hyperlipidemia	44.3	42.0	51.0	64.5	75.9	74.4	67.6	0.406	0.411
Liver Disease	3.9	5.7	9.0	53.3	96.2	60.0	93.0	0.404	0.316
Arrhythmia	16.8	35.8	53.6	72.7	70.3	89.4	44.4	0.308	0.045
Chronic Obstructive Pulmonary Disease	10.1	19.3	29.4	53.8	84.5	52.5	55.7	0.26	0.161
Congenital Heart Disease	0.8	2.6	3.6	33.3	97.7	66.7	86.5	0.139	0.0979
Substance Abuse	3.4	0.8	4.1	7.7	99.5	15.4	96.3	0.107	0.104
Congestive Heart Failure	41.2	30.7	47.7	35.6	72.8	33.3	96.6	0.077	0.076
Smoking	27.6	2.6	22.9	8.4	99.6	46.7	86.1	0.003	0.345
Average	23.1	25.3	28.6	59.6	82.1	59.8	77.6	0.391	0.225