Template Matching for Auditing Hospital Cost and Quality

Jeffrey H. Silber, Paul R. Rosenbaum, Richard N. Ross, Justin M. Ludwig, Wei Wang, Bijan A. Niknam, Nabanita Mukherjee, Philip A. Saynisch, Orit Even-Shoshan, Rachel R. Kelz, and Lee A. Fleisher

Objective. Develop an improved method for auditing hospital cost and quality.

Data Sources/Setting. Medicare claims in general, gynecologic and urologic surgery, and orthopedics from Illinois, Texas, and New York between 2004 and 2006.

Study Design. A template of 300 representative patients was constructed and then used to match 300 patients at hospitals that had a minimum of 500 patients over a 3-year study period.

Data Collection/Extraction Methods. From each of 217 hospitals we chose 300 patients most resembling the template using multivariate matching.

Principal Findings. The matching algorithm found close matches on procedures and patient characteristics, far more balanced than measured covariates would be in a randomized clinical trial. These matched samples displayed little to no differences across hospitals in common patient characteristics yet found large and statistically significant hospital variation in mortality, complications, failure-to-rescue, readmissions, length of stay, ICU days, cost, and surgical procedure length. Similar patients at different hospitals had substantially different outcomes.

Conclusion. The template-matched sample can produce fair, directly standardized audits that evaluate hospitals on patients with similar characteristics, thereby making benchmarking more believable. Through examining matched samples of individual patients, administrators can better detect poor performance at their hospitals and better understand why these problems are occurring.

Key Words. Quality of care, outcomes research, health care research, cost

There are reasons to audit hospitals with respect to cost and quality. An insurer, such as Medicare, Medicaid, or a private insurance company, wants to ensure efficient and safe practices. A hospital's Chief Medical Officer wants to benchmark efficiency and quality in comparison to performance at other hospitals.

Since patients are not randomly allocated to hospitals, hospitals understandably fear unfair comparisons in which differing outcomes are attributed to poor quality when in fact they reflect a sicker patient population (Berwick and Wald 1990). Hospitals are commonly compared by indirect standardization (Kitagawa 1955; Silber, Rosenbaum, and Ross 1995; Fleiss, Levin, and Paik 2003; Iezzoni 2012). Indirect standardization describes how a hospital's observed patient outcomes compare with how those same patients may have fared at the typical or average hospital. Such an analysis allows the substrate of information used for comparisons (observable patient case mix and severity) to vary across hospitals—that is, the patients at each hospital may be quite different. However, through model adjustments, analysts can attempt to make important comparisons regarding outcomes and quality.

Unlike indirect methods, we will develop quality audits using multivariate template matching—a form of direct standardization (Fleiss, Levin, and Paik 2003). Direct standardization considers a fixed patient population and asks how different hospitals would perform with the same or similar patients. Using this method, we select similar patients at different hospitals so that comparison of hospital outcomes and processes are fair and easily interpretable. Furthermore, because we are constructing actual matched samples of patients at each hospital, template audits produce actionable interventions to better understand why specific hospitals fared better or worse because individual patients, or groups of patients, in the template can be closely examined. Our

Address correspondence to Dr. Jeffrey H. Silber, Center for Outcomes Research, The Children's Hospital of Philadelphia, 3535 Market Street, Suite 1029, Philadelphia, PA 19104; e-mail: silberj@wharton.upenn.edu. Jeffrey H. Silber, M.D., Ph.D., is also with The Department of Pediatrics, The University of Pennsylvania School of Medicine, Department of Anesthesiology and Critical Care, The University of Pennsylvania School of Medicine, Department of Health Care Management, The Wharton School, The University of Pennsylvania, The Leonard Davis Institute of Health Economics, The University of Pennsylvania; Philadelphia, PA. Paul R. Rosenbaum, Ph.D., is with The Leonard Davis Institute of Health Economics, The University of Pennsylvania, Philadelphia, PA and the Department of Statistics, The Wharton School, The University of Pennsylvania, Philadelphia, PA. Richard N. Ross, M.S., Justin M. Ludwig, M.A., Wei Wang, Ph.D., Bijan A. Niknam, B.S., Nabanita Mukherjee, Ph.D., and Philip A. Saynisch, A.B., are with the Center for Outcomes Research, The Children's Hospital of Philadelphia, Philadelphia, PA. Orit Even-Shoshan, M.S., is also with the Center for Outcomes Research, The Children's Hospital of Philadelphia, The Leonard Davis Institute of Health Economics, The University of Pennsylvania, Philadelphia, PA. Rachel R. Kelz, M.D., is with the Department of Surgery, The University of Pennsylvania School of Medicine, Philadelphia, PA. Lee A. Fleisher, M.D., is also with the Department of Anesthesiology and Critical Care, The University of Pennsylvania School of Medicine, The Leonard Davis Institute of Health Economics, The University of Pennsylvania, Philadelphia, PA.

goal is to make hospital audits more credible, and useful, to the Chief Medical Officer.

METHODS

Conceptual Model and Statistical Methods

It is easiest to understand the template audit as analogous to a classroom examination. A fair exam asks each student the same or similar questions. A fair audit contrasts hospital performance on similar patients, patients undergoing similar procedures with the same or similar comorbid conditions.

The template, the exam, is a collection of 300 patients undergoing procedures performed at most hospitals. The template excludes rare types of procedures and patients, types that occur with a frequency substantially less than 1/300, and procedures that are performed only at some hospitals. Within each hospital, 300 patients are individually paired to the 300 patients in the template. If the first patient in the template is a 68-year-old woman with a prior heart attack undergoing hip surgery, then a similar patient is found at the hospital. This process of “matching” hospital patients to the template’s patients is accomplished using multivariate matching (Rosenbaum 2010b). In the end, we will evaluate each hospital using their 300 matched patients selected to resemble the template. Every hospital takes the same exam. The hospital sees outcomes and costs for 300 of its own patients together with summaries of outcomes for 300 similar patients at hundreds of other hospitals. Because these are 300 real patients at the hospital, not theoretical coefficients in a model, the hospital may examine its own patients with poor outcomes or excessive costs in as much detail as they care to consider. Later, in Table 2, we present one such audit for one hospital, comparing its performance with 300 patients to the performance of 216 other hospitals with 300 very similar patients.

Constructing the Examination

For this research we examined the cost and quality of hospitals treating Medicare patients admitted for orthopedic and general surgery (including some urological and gynecological procedures often performed by general surgeons) throughout New York, Illinois, and Texas using the ICD-9-CM principal procedure codes found in the claims. We chose these states because their Medicare patients had a relatively low rate of managed care as compared with other states, and the states were geographically diverse. For these states we

obtained the Medicare Part A, Part B, and Outpatient files for the years 2004–2006, and merged this file to the Social Security denominator file to determine dates of death and other demographic information.

We chose a template size of 300 because of practical size constraints and power considerations, though templates of various sizes can be constructed, depending on the purpose of the audit. Choosing a template of 300 prompted us to select hospitals that performed at least 500 cases over the 3 years of the dataset to ensure adequate matching ratios to achieve good matches at each hospital. Our study dataset had 217 usable hospitals. Using a template with 300 patients, we will wish to compare an outcome rate at a single hospital to the remaining 216 hospitals. To provide a rough sense of the statistical power for such a template size, we utilize the method of Miettinen (Miettinen 1969). For example, using a two-tailed type-I error of 5 percent, a hospital with an 8 percent 30-day mortality rate compared to a 4 percent rate for the remaining hospitals could be detected with above 90 percent power, and comparing a hospital with a 50 percent 30-day complication rate to the remaining hospitals with a 40 percent rate would obtain a power well above 90 percent. We provide these power calculations only as a rough guide for the reader. The excellent power stems in part from the fact that each of 300 patients at one hospital is compared to 216 similar patients at 216 other hospitals, producing a comparison group of size $216 \times 300 = 64,800$ patients.

To construct the template, we generated 500 random samples of size 300, comprised of 100 general, gynecologic, and urologic surgery of the type often performed by general surgeons and 200 orthopedic surgery cases (reflecting the fact that orthopedic surgery was more common in our hospitals than general surgery). We wanted the template to look like the population in terms of age, sex, procedure, procedure cluster, predicted procedure time, probability of death score, 31 comorbidities, transfer-in status, and emergency status, so we picked the best of the 500 random samples. The best sample had means that were “closest” to the population means for the variables just mentioned. “Closest” was defined using the Mahalanobis distance between the means—the most standard of multivariate distance measures—and it is essentially the sum of the squares of the differences in means in units of the standard deviation with an adjustment that reduces the redundancy of variables that are highly correlated (see Rubin 1979, 1980).

For 300 patients to be close to the population, extremely unusual patients who occur with a frequency far less than $1/300$ would have to be absent from the 300. The winning template was defined as our “exam” for this analysis, representing a patient population that likely would achieve good

matches across many hospitals (a full description of the template is found in the Appendix).

It is important to realize that our decision to find typical surgical patients is just one approach to auditing hospitals. Instead, we may have wanted to concentrate the audit on a different set of patients. We could have chosen a mix of common medical conditions or a mix of medical or surgical patients. We could have weighted the exam to select more difficult patients for the template. Our intent in this report is to provide an example that is useful and illustrates this new methodology. Templates should be constructed with the purpose of the audit in mind.

Administering the Examination

The process of giving the exam (performing the audit) involves the selection of 300 patients at each hospital who closely resemble the 300 patients in the template. We call this process “template matching” because at each hospital we choose 300 patients at that hospital that most reflect the 300 patients in the template: every hospital takes the same exam. To match 300 patients to the template requires some pool of patients at each hospital. As our template includes 100 general, gynecologic and urologic surgery patients and 200 orthopedic patients, we only analyzed hospitals that saw at least 200 general, gynecologic or urologic surgery patients and 300 orthopedic surgery patients. We used a 1:2 ratio for general surgery because there were so many different procedure types that we thought we may have encountered a harder time matching these patients than in orthopedic surgery, with considerably fewer procedures. Therefore, in orthopedics, we only required a 2:3 matching pool.

The matching was accomplished using Medicare claims from 2004 to 2006 (approximately 3 years of data, less 3 months to allow for a look-back period when defining comorbidities). We performed our matches using R MIP Match (R Development Core Team 2012; Zubizarreta 2012; Zubizarreta, Cerda, and Rosenbaum 2013) and specified the following algorithm for selecting matches to the overall template: Match exactly on principal procedure whenever possible; if not possible, attempt to match within a procedure cluster (a clinical group of procedures that resemble the index procedure; for example, right versus left hemicolectomy are in the same procedure cluster; see Appendix); if not possible, match within a set of anatomically similar procedure clusters, and finally, if not possible, match to other general surgical procedures. Inside each hierarchical category, we choose a match based on the minimized medical distance between patients, where medical distance is

defined through the Mahalanobis distance, similar to what was described above for choosing the template. Details concerning the elements of the Mahalanobis distance are provided in the Appendix.

To improve the quality of the matches between the template and the specific hospital, we utilized fine balance (Rosenbaum, Ross, and Silber 2007; Silber et al. 2007b, 2012; Rosenbaum 2010a; Yang et al. 2012) within general surgical and orthopedic patients. Fine balance says that if one must tolerate a mismatch on, say, CHF for one patient, then this mismatch must be counterbalanced by a mismatch for CHF in the opposite direction, so the number of patients with CHF at the hospital equals the number in the template. Fine balance ensured that if the template had, say 25 percent CHF cases for its 100 general surgical cases and 35 percent CHF for its orthopedic cases, each hospital also provided a 25 percent rate of CHF for its general surgical and 35 percent rate for its orthopedic surgery cases whenever possible, without absolutely requiring exact matches on CHF for each and every patient in the hospital with respect to the template, though the algorithm also preferred to do that as often as possible via minimizing the Mahalanobis distance function.

Grading the Fairness of the Examination

Ideally, for a fair exam, every hospital would have been tested on exactly the same 300 patients as every other hospital. Of course, this is not possible—yet we can evaluate the fairness of the examination or audit by observing how similar the characteristics of the matched patients are across hospitals. If each hospital's 300 patients display similar patient characteristics, then we will define this as a fair exam. For each hospital we can formally test whether the matched patients at their hospital are similar to the template. We do this in two steps. First, we tested whether our matching algorithm found similar rates of ICD-9-CM matches on principal procedures for the major groups of procedures (in our case six groups included 85 percent of patients in the template). We compared the rate of the important surgical clusters individually for the matched patients in a particular hospital versus the template patients using Fisher's Exact Test, and excluded hospitals where a Bonferroni-corrected p -value was significant at the .05 level, as these hospitals were unable to be matched adequately on the important surgical clusters. Next, we also formally tested whether the hospital's matched set was similar to the template using the "cross-match" test (Rosenbaum 2005; Heller et al. 2010a; Heller, Rosenbaum, and Small 2010b; Silber et al. 2013a). The cross-match test determines whether a hospital could be distinguished from the template based on the

patients in the hospital's matched set. The cross-match test will pair patients on the basis of their characteristics, then asks whether these pairs separate the hospital's patients from the template's patients. The cross-match test provides both the number of pairs that displayed significant differences from the template and the matched hospital set, and an associated distribution free p -value. A hospital whose matched patients were not significantly different from the template would display an insignificant p -value and have few patients who "failed" the cross-match test.

Is the Template "Representative" of the Patients at an Individual Hospital?

The cross-match test can assess whether a template differs from the patients typically treated at an individual hospital. For each hospital we randomly select 300 patients (a representative sample), pairing the 600 = 300 template + 300 hospital patients using patient characteristics. Using the cross-match test as above, we now ask how many cross-matches occur between the template and the hospital sample. If the template resembles the hospital's patients, there will be many cross-matches and Upsilon will be high; otherwise, there will be fewer cross-matches and Upsilon will be low. The template differs significantly ($p \leq .05$) from the hospital's patients if there are less than 134 cross-matches, or equivalently an Upsilon statistic less than 0.45.

Grading the Hospitals

Matching is done first, to create a fair exam, without viewing outcomes. Once the hospitals matches have been deemed fair, through examining the quality of the matches and using the cross-match test, the matching process stops, and examination of patient outcomes, costs, or processes at the hospital can begin. This two-step process prevents multiple analyses to find an analysis that makes outcomes look good—it prevents editing the exam after the student has answered the questions (Rubin 2007, 2008).

Here, we examine in-hospital and 30-day mortality; in-hospital and 30-day complications; in-hospital and 30-day failure-to-rescue (Silber et al. 1992, 2007a); readmissions within 30 days of discharge; resource utilization based costs, both in-hospital and 30 day (Silber et al. 2013b); length of stay; the percent of patients that were in the ICU and the length of ICU stay for those in the ICU; and a potential process metric available in the dataset, operative procedure length as defined through the anesthesia bill (Silber et al. 2007c,d, 2011, 2013b). Readmission is all cause readmission for this example, and complications are

defined as in previous work (Silber et al. 2007a, 2012) comprising 38 common complications that occur postoperatively. Each hospital audit would include a comparison of the outcomes at that hospital to the other 216 hospitals in the study. The audit will also include postmatch adjustments (Cochran and Rubin 1973; Rubin 1979; Silber et al. 2005) that are expected to be generally similar to the “stratified” rates since we have already performed extensive matching to produce similar patients at each hospital. The stratified analysis will be performed using a Mantel–Haenszel test for discrete variables (clustering on the template patient) and a stratified Wilcoxon rank sum test (Lehmann 2006). The adjustment model for each discrete outcome used conditional logistic regression clustering on the template matched patient adjusting for probability of death score, predicted procedure time, and an indicator for emergency admission. For each continuous outcome, we used robust regression (Huber 1981; Hampel et al. 1986), adjusting for the same variables and an indicator variable describing the template patient for which the observed patient was

Figure 1: Template Matching Conceptual Model. (A) We select 300 patents from the population to form a template (shapes represent different patients in the template). We chose a template that had patients with common characteristics across a large number of hospitals. From each of 217 hospitals we select 300 patients who are most like the template. We then have a sample of $300 \times 217 = 65,100$ patients. (B) Our goal is to find matches such that each hospital will have 300 patients who will be similar to the template, and therefore similar to each other’s 300 patients. We depict this as similar hospital patients in each of the hospital matched samples. Overall, 300 patients at one hospital are very similar to the 300 patients at any other hospital. We formally test whether patients at one hospital are different from the template using the cross-match test (Rosenbaum 2005; Heller et al. 2010a) and examine whether the probability of death or rates of principal procedures are different between the hospital and the template using the paired Wilcoxon rank sum test (Lehmann 2006) or the Mantel–Haenszel test, respectively. (C) For each patient in the template there are 217 representations, one at each hospital. In general, each patient from the template has a close representation at any hospital. (D) represents the fine balancing that allows for both similar patient procedures across each template, but also similar covariates such as age, probability of death, rate of diabetes, and rate of CHF, to name a few. With such similarities, it becomes more difficult for a hospital CMO to suggest their patients are somehow different from the template

matched to (Rubin 1979). To determine the potential impact of failure to match for some unobserved covariate on outcomes results, we performed a sensitivity analysis (Rosenbaum 1988, 2002; Rosenbaum and Silber 2009).

Learning from the Examination

In examining hospitals, learning from an exam is probably far more important than the grade itself. Each hospital will have 300 patients on which outcomes and process can be closely evaluated and benchmarked across the other 216 hospitals in the study. Such results can be examined as a group or analyzed separately. Hospitals can examine why they fared poorly on the audit, and why they have an immediate reference population for each patient in their audit.

In Figure 1 we describe the conceptual model of the template matched audit. We start with all patients (with some shown at the bottom of Figure 1A) and select a template as described above. We have depicted each procedure in the template with a different shape. We have shown four shapes in the template, a pentagon, star, triangle, and circle representing 4 of 300 patients in

the template. From each hospital we select 300 patients with characteristics as close as possible to the template patients. Note, we display these shapes as close but not exact copies across hospitals. Patients with the same procedure but different comorbidities or ages are depicted with different size shapes. In Figure 1B we display the 217 hospitals, each with similar patients to the template. We can then examine outcomes of each hospital's 300 patients to "grade" the hospitals. In Figure 1C we display that each patient in the template has 217 representations (one at each hospital). We can then examine the outcomes of each of these 217 patients, or groups of patients, to understand the difficulty of these patients to better benchmark the practice of individual hospitals (as will be described). Figure 1D represents the fine balancing that allows for both similar patient procedures across each template, but also similar covariates balancing simultaneously numerous variables such as age, probability of death, rate of diabetes, and rate of CHF, to name a few.

Comparing Template Matching to Traditional Regression Methods

To better describe how template matching may differ from standard risk adjustment using regression, we utilized a risk model for mortality, and compared rankings between methods. To accomplish this, we performed a number of comparisons: (1) Rematched the original template to each hospital but utilized the risk score as the only matching variable. The template match compares patients who have the same risk for the same reason—for example, both patients have CHF—but model-based risk adjustment compares patients that a model's risk score says have the same risk of death, perhaps for different reasons—for example, one has CHF, the other diabetes. We then compared the similarity of patient characteristics across hospitals and compared the rank ordering of hospitals based on various mortality based outcomes; (2) compute an O/E measure based on all patients at each hospital, rank hospitals based on this and compare to the Template matched results; and finally (3) compute an O/E ratio for each hospital using only the 300 matched patients based on the 300 risk-matched patients at each hospital.

RESULTS

Hospital and Patient Populations

We required that hospitals had at least 200 general, gynecologic and urologic surgery cases and 300 orthopedic cases over the 3-year period for us to include

them in the analysis. Of 514 acute care facilities within New York, Illinois, and Texas that participate in Medicare's Surgical Care Improvement Project, 241 met this size requirement. These 241 hospitals comprised 331,628 orthopedic and general surgery cases, or 72.9 percent of the 455,010 procedures performed at all 514 hospitals. Of these hospitals, we had complete data on 228 hospitals, and of these 228 we could successfully match 217 hospitals to the template, meaning that four hospitals failed the cross-match test despite being of adequate size and having full billing data and seven displayed a significant difference in the Bonferroni adjusted Fisher's Exact Test on the rates of the six main procedure clusters that comprise 85 percent of procedures in the template. We therefore report on 217 of 228 hospitals or 95 percent of the hospitals that had adequate size and adequate data quality. Compared with the initial set of 241 hospitals, these 217 hospitals contained 308,882 of the 331,628 procedures performed, or 93.1 percent. Compared with all 514 hospitals in the three states, we used 67.9 percent of all procedures performed. In an actual audit, hospitals would always be able to identify if the matching quality was adequate because they would be able to observe their patients and the template, as well as the other hospitals' patients and therefore would clearly be able to scrutinize the fairness of comparisons. Furthermore, we could have increased distribution of procedure type to include more hospitals if desired, but this was not the intent of this report. We could also have used a different matching design, called full matching, that would create matched sets of variable size, perhaps allowing some of the 11 hospitals who failed our matching criteria to be included; however, this would require a weighted analysis (Hansen 2004, 2007; Hansen and Klopfer 2006; Rosenbaum 1991).

Note, For the 217 study hospitals, the mean bed size was 384 with 63.4 percent greater than 250 beds and 4.1 percent less than 100 beds. Nonteaching hospitals represented 67.3 percent of the study hospitals, and major or very major teaching hospitals (with resident to bed ratios greater than 0.25) represented 7.2 percent of the 217 hospitals.

Examining the Quality of the Matches Using the Full Template and the Risk-Score Match

In Table 1 we examine how similar the hospital's matched samples were across hospitals and formally test the overall variation of patient characteristics and outcomes across hospitals using the Kruskal–Wallis test, a nonparametric version of the one-way ANOVA test (Kruskal and Wallis 1952) for each continuous variable of interest and the Pearson Chi-square test for each

Table 1: Assessing Whether Patient Covariates and Outcomes Vary Significantly across Hospitals ($N = 217$ hospitals). For both the full template match based on using all matching variables (Template) and a match to the same template using only Risk Score (Risk), we compare the variation among individual hospitals in patient covariates and outcomes to the variation that would have been expected had patients been randomly assigned to hospitals. For the chi-square statistics, a chi-square/degrees of freedom ($= 216$ for 217 hospitals) that is greater than 1 suggests more variation than random, and less than 1 suggests less variation than random. For continuous variables (except probability of death and Length of Stay), we display the Hodges–Lehmann estimates* and report the statistic and p -value using the Kruskal–Wallis test. Note that patient variables are very similar across hospitals, whereas hospital outcomes are not. However, for the Risk match, all displayed variables significantly vary across hospitals except the matching variable (risk score). For each variable, the template match is shown first (left side), then the risk score-only match (right side)

Patient Covariates	Percentile Range for Template Matches (Means or Hodges–Lehmann Estimates)					Percentile Range for Risk-Only Matches (Means or Hodges–Lehmann Estimates)					Chi-Square Statistic/df	p Value
	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)		
Age (years)	76.3	76.5	76.7	76.9	77.1	76.7	77.0	77.4	77.9	78.2	2.4519	<.0001
Gender (% male)	33.0	33.3	34.3	35.0	35.0	29.7	31.7	34.0	36.7	38.3	1.9811	<.0001
Probability of 30-day death	0.031	0.032	0.034	0.035	0.036	0.0326	0.0327	0.0329	0.0331	0.0333	0.0416	1.0000
Predicted procedure time (min)	135.1	135.7	136.6	137.1	137.4	132.7	133.8	135.2	137.5	139.2	4.9084	<.0001
Emergency admission	32.0	32.0	33.0	34.0	35.3	27.3	29.7	33.3	36.0	38.0	4.3296	<.0001

continued

Table 1: Continued

Patient Covariates	Percentile Range for Template Matches (Means or Hodges-Lehmann Estimates)						Percentile Range for Risk-Only Matches (Means or Hodges-Lehmann Estimates)						Chi-Square Statistic/df	p-value	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)	Chi-Square Statistic/df	p-Value
	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)		Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)										
Transfer-in†	0.0	0.0	0.7	1.3	1.7		0.0	0.0	0.3	1.7	3.3		2.1684	<.0001	0.0	0.0	0.3	1.7	3.3	11.0931	<.0001
Comorbidities																					
CHF	18.7	19.0	20.0	20.0	20.0		15.3	16.7	18.7	21.0	23.0		0.1313	1.0000	15.3	16.7	18.7	21.0	23.0	2.4450	<.0001
Past	23.7	24.0	25.0	26.0	26.0		20.7	23.0	25.7	28.3	31.0		0.2776	1.0000	20.7	23.0	25.7	28.3	31.0	2.3331	<.0001
arrhythmia																					
Past	5.0	5.7	6.3	6.3	6.3		3.7	5.0	6.0	7.7	9.0		0.2741	1.0000	3.7	5.0	6.0	7.7	9.0	4.0755	<.0001
myocardial																					
infarction																					
Angina	1.7	2.0	2.3	3.0	3.7		1.0	1.7	2.7	4.3	5.7		1.2349	.0162	1.0	1.7	2.7	4.3	5.7	5.3689	<.0001
Diabetes	22.7	23.3	23.7	23.7	24.3		20.3	22.0	24.7	28.0	30.3		0.1432	1.0000	20.3	22.0	24.7	28.0	30.3	4.2875	<.0001
Renal	6.3	7.0	8.0	8.0	8.0		4.7	5.7	6.7	8.3	10.0		0.3504	1.0000	4.7	5.7	6.7	8.3	10.0	2.5498	<.0001
dysfunction																					
COPD	19.3	20.0	20.3	20.3	21.0		15.0	17.3	20.0	23.7	26.0		0.1587	1.0000	15.0	17.3	20.0	23.7	26.0	5.2299	<.0001
Asthma	5.0	5.3	6.0	6.7	7.3		5.0	5.7	7.0	8.3	10.0		0.4474	1.0000	5.0	5.7	7.0	8.3	10.0	2.0764	<.0001

continued

Table 1: Continued

Outcomes	Percentile Range for Template Matches (Means or Hodges–Lehmann Estimates)					Percentile Range for Risk-Only Matches (Means or Hodges–Lehmann Estimates)					Chi-Square Statistic/df	p-Value		
	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)	Lower Eighth (12.5th)	Lower Quartile (25th)	Median	Upper Quartile (75th)	Upper Eighth (87.5th)				
Mortality														
Inpatient%	1.0	1.3	1.7	2.3	2.7	1.1271	.0967	1.0	1.3	1.7	2.3	2.7	1.0616	.2549
30-day%	2.0	2.3	3.0	4.0	4.7	1.3038	.0018	2.0	2.7	3.3	4.0	4.7	1.1637	.0497
Complications														
Inpatient%	42.0	45.0	49.3	54.3	59.7	6.8264	<.0001	41.0	45.7	50.0	55.7	59.7	7.1076	<.0001
30-day%	53.7	57.7	63.0	68.0	71.0	7.2533	<.0001	54.7	57.7	63.7	69.7	72.7	8.2178	<.0001
Failure-to-rescue														
Inpatient%	1.7	2.4	3.4	4.7	5.5	1.3421	.0006	1.7	2.4	3.5	4.8	5.6	1.1907	.0287
30-day%	2.8	3.6	5.0	6.4	7.8	1.6138	<.0001	3.0	4.0	5.0	6.4	7.4	1.4005	<.0001
Readmissions														
30-day%	7.3	8.3	9.7	11.3	12.7	1.8222	<.0001	7.7	8.7	10.0	11.7	12.7	1.6760	<.0001
LOS†	4.8	5.1	5.5	5.8	6.1	9.9954	<.0001	4.8	5.0	5.5	5.8	6.1	9.5143	<.0001
Patients in ICU (%)	9.0	11.7	17.0	24.3	29.3	25.8929	<.0001	8.3	11.7	17.3	24.0	30.3	27.5258	<.0001
Days in ICU, if sent	3.5	4.4	5.3	6.2	7.3	6.6975	<.0001	3.4	4.1	5.0	6.1	7.1	7.3995	<.0001
Total cost (\$K)														
Inpatient \$	11.3	12.1	13.0	13.9	14.8	17.0511	<.0001	11.4	12.1	13.1	14.0	14.9	17.3094	<.0001
30-day \$	12.3	13.1	14.0	15.1	16.3	14.7172	<.0001	12.2	13.1	14.2	15.2	16.1	15.1992	<.0001
Procedure time (min)	117.0	123.7	133.5	147.0	157.5	41.4650	<.0001	117	123.8	134.2	145.5	157.5	42.3730	<.0001

Note. †The Hodges–Lehmann estimate is the location estimate derived from and compatible with the Wilcoxon signed-rank statistic.

*Transfer-in was significant but represented only a range from zero patients (lowest hospital) to ten patients (highest hospital).

†LOS (length of stay) hospital means were calculated using trimmed means, excluding 2.5% of patients from each extreme.

Correction added on 22 August 2014 after first online publication on 3 March 2014: Some data on page 13 were move on page 14 and column head “Patient Covariates” has been changed to “Outcomes”.

binary variable. The test is simply a benchmark, comparing the similarity of hospitals in our template match to the similarity of hospitals if patients had been randomly assigned to hospitals. The chi-square statistic for the Kruskal–Wallis test divided by its degrees of freedom is also reported. In a completely randomized clinical trial—that is, random assignment of patients to hospitals—the Kruskal–Wallis chi-square divided by degrees of freedom would have expectation = 1; it would tend to be larger than 1 if patients differed more than expected by random assignment, and it would have expectation less than 1 if patients differ across hospitals less than expected under random assignment. Arguably, the exam looks fair if this chi-square ratio is about 1 or less than 1.

For the full template-based match (the left side of the table) we see that patient covariates across hospitals are very similar, far more similar than if the patients were randomly assigned to hospitals, as the chi-square ratios are generally far below 1 and *p*-values are near 1.0. For major patient characteristics and hospital outcomes, Table 1 presents the hospital values for half the entire distribution (the median), then divided by quarters and then half again (12.5th and 87.5th percentiles). Template matching made groups of patients at different hospitals look far more similar in terms of the variables in Table 1 than would have been expected by randomly assigning patients to hospitals. The procedure matching algorithm was very successful: of 65,100 patients from the 217 hospitals, 63,939, or 98.2 percent, were exactly matched to their respective template patient on ICD-9-CM principal procedure. It is possible, of course, that patients differ in ways not recorded in Table 1. Compare the similarity in patient characteristics for the template match to the risk score-based matching that was used for the right-hand side of Table 1. Like a typical regression approach, the Risk score-only match compares patients that the model judges to have similar risks of death, even if they have similar risks of death from different patterns of comorbidities. As seen on the right in Table 1, matching for risk of death makes patient groups similar in terms of risk of death, but dissimilar in terms of the patient characteristics that comprise that risk of death. Except for the risk score variable itself, patient characteristics were very different across hospitals in the risk score match, differing by far more than would be expected in a clinical trial and by far more than when using template matching.

The bottom of Table 1 looks at outcomes asking if similar patients at different hospitals have similar outcomes. In both the Template Match and the Risk Match, unlike patient characteristics, hospital outcomes varied greatly, and significantly, across institutions. Outcome rates differ among hospitals by

Figure 2: Hospital Outcomes. We display the outcomes for 217 hospitals, each hospital having 300 template-matched patients. For each outcome we provide density plots with an associated box plot providing 5th, 25th, 50th, 75th, and 95th percentile markers, as well as a depiction of outliers. We also denote individual Hospital A (examined in Table 2) as a large solid dot on the plots. Displayed are: (A) in-hospital mortality; (B) 30-day mortality; (C) in-hospital failure-to-rescue rate; (D) 30-day failure-to-rescue rate; (E) in-hospital complication rate; (F) 30-day complication rate; (G) percentage of patients using the ICU; (H) ICU LOS in patients using the ICU; (I) 30-day readmission rate; (J) length of stay (2.5 percent trimmed mean); (K) total costs of the index admission (Hodges–Lehmann estimates); (L) total cost of the index admission, plus total costs 30 days after discharge from the index admission (Hodges–Lehmann estimates)

far more than would be expected if hospitals provided equivalent care and patients were randomly assigned to hospitals. On covariates, patients look more similar than by random assignment, but on outcomes they look quite different. This is also described in Figure 2, where we examine each of the reported outcomes across the 217 hospitals with probability density plots, where the area under each distribution sums to a probability of 1.0.

We also examined the correlation between the Template match hospital mortality rankings and the rankings based on various risk based approaches. Using all patients at each hospital, we compute an O/E ratio and rank order the hospitals. The Spearman correlation between this measure and the Template matching mortality ranking was 0.69, 95 percent CI (0.62, 0.76), $p < .0001$. We next examined the Spearman correlation between the O/E ratio based on the risk match ($N = 300$ patients from each hospital matched to the template) and the template match, $r = 0.54$, 95 percent CI (0.44, 0.63), $p < .0001$. Evidently, comparing patients with the same risk for the same reason differed from comparing patients with the same risk, perhaps for different reasons.

Finally, we examined the correlation between the quality of the matches as defined by Upsilon and actual mortality based on the template. We found the correlation was -0.09 , 95 percent CI (-0.22 , 0.05), $p = .192$. The variation in mortality among hospitals after matching is not predicted from the closeness of the match to the template, perhaps because most matches were quite close. When we examined the correlation between the “Representative” Upsilon test derived from comparing the template to each hospital’s random

Figure 2: *Continued*

Table 2: Understanding an Individual Hospital's Outcomes: A Hospital Profile to Illustrate the Audit Results.

Outcomes (Percent Unless Otherwise Noted)	All Patients			Orthopedic Surgery Patients			General Surgery Patients		
	Hospital A (N = 300)		216 Other Hospitals (N = 64,800)	Hospital A (N = 200)		216 Other Hospitals (N = 43,200)	Hospital A (N = 100)		216 Other Hospitals (N = 21,600)
	% or Mean	Percentile		% or Mean	Percentile		% or Mean	Percentile	
Mortality									
Inpatient [†] %	3.3 [†]	95	1.7	2.5	95	1.0	5.0	81	3.2
30-day [‡] %	6.0*** ^{††}	98	3.2	5.5*** ^{††}	97	2.6	7.0	88	4.3
Complications									
Inpatient [†] %	53.7	70	50.3	48.5	61	47.4	64.0	88	56.0
30-day [‡] %	59.7	35	62.9	55.5** [†]	24	63.0	68.0	83	62.7
FTR									
Inpatient [†] %	6.2	97	3.6	5.2	91	2.2	7.8	73	5.8
30-day [‡] %	10.1** ^{††}	93	5.2	9.9** [†]	97	4.3	10.3	82	6.9
Readmissions									
30-day [‡] %	9.0	36	9.9	9.5	52	9.5	8.0	22	10.8
Total cost (\$k)									
Inpatient	13.6	66	13.1	10.9***** ^{†††}	25	11.9	27.5***** ^{††††}	98	18.0
30-day	15.2	76	14.2	11.8***** ^{††}	29	12.7	28.7***** ^{††††}	97	19.4
Overall length of stay	5.8	75	6.0	4.6	42	5.4	9.0 ^{††}	96	7.2
% of Patients sent to ICU	28.7***** ^{††††}	84	19.3	12.5	66	12.1	61.0***** ^{††††}	97	33.6
Days in ICU, if sent to ICU ^{†‡}	7.4***** ^{††††}	88	5.3	4.9	68	4.3	8.5***** ^{††††}	89	6.1
Procedure time (minutes)	150.8***** ^{††††}	80	135.8	146.2***** ^{††††}	71	135.0	165.0***** ^{††††}	88	138.8

continued

Table 2. Continued

Balance of Matched Variables	All Patients		Orthopedic Surgery Patients		General Surgery Patients	
	Hospital A (N = 300)	216 Other Hospitals (N = 64,800)	Hospital A (N = 200)	216 Other Hospitals (N = 43,200)	Hospital A (N = 700)	216 Other Hospitals (N = 21,600)
	% or Mean		% or Mean		% or Mean	
Age (mean, years)	77.0	76.7	77.5	77.3	76.0	75.5
Age (% >80 years)	34.3	32.7	36.5	35.8	30.0	26.7
Gender (% male)	34.7	34.2	33.5	33.1	37.0	36.5
Procedure time (predicted)	137.1	136.1	134.1	137.1	141.1	138.5
Probability of 30-day death	0.033	0.034	0.030	0.030	0.038	0.042
Emergency admission (%)	32.0	33.1	29.0	29.4	38.0	40.5
Transfer-in (%)	1.7	0.8	1.5	0.9	2.0	0.7
History of CHF (%)	20.0	19.6	18.0	17.7	24.0	23.3
History of arrhythmia (%)	26.0	24.9	25.5	25.3	27.0	24.3
History of MI (%)	6.3	5.9	5.5	5.2	8.0	7.4
History of angina (%)	2.3	2.6	3.0	2.8	1.0	2.2
History of diabetes (%)	23.7	23.6	22.5	22.5	26.0	25.8
History renal disease (%)	8.0	7.4	7.0	6.5	10.0	9.3
History of COPD (%)	20.3	20.2	18.5	18.4	24.0	23.7
History of asthma (%)	7.3	6.1	7.0	6.6	8.0	7.4

Note. Letters denote p-values for comparing Hospital A to the 216 other hospitals. Percentile ranks (1 = best, 100 = worst). Conditional logit and m-estimation p-values account for matched differences in probability of death, predicted procedure time, and emergency. Length of Stay means were calculated using trimmed means, excluding 2.5% of patients from each extreme. Total costs are reported using Hodges-Lehmann estimates. Mantel-Haenszel (binary variables) or stratified Wilcoxon rank-sum (continuous variables) p-value key: * $p < .05$; ** $p < .01$; *** $p < .001$; **** $p < .0001$.

Conditional logit (binary variables) or m-estimation (continuous variables) p-value key: † $p < 0.05$; ‡ $p < .01$; +++ $p < .001$; ++++ $p < .0001$.

*Tests for differences in days in ICU, if sent to ICU, are not paired due to the nature of the variable.

sample of 300 patients, the Spearman correlation between Upsilon and mortality was 0.19, 95 percent CI (0.06, 0.32), $p = .005$.

Examining the Individual Hospitals—The Hospital Audit

We can now proceed to audit an individual hospital—the main goal of this report. We examine Hospital A on the outcome and process metrics of interest, and compare Hospital A to the rest of the patients matched to the template at each of the 216 hospitals in the dataset, since all hospitals were matched to the same initial template.

In Figure 2 Hospital A is denoted with a large solid dot in all graphs. In Table 2 we observe that Hospital A displayed a higher in-hospital death rate (3.3 percent) than the grouped rate of the other 216 control hospitals (1.7 percent), or a rank ordering of 95th percentile, and a similar poor performance in 30-day mortality with a 6.0 percent death rate versus controls of 3.2 percent, and a 98th percentile score. Without any post-match adjustment, the death rates for 30-day mortality did reach statistical significance ($p = .004$). With post-match adjustment for the probability of death, emergency room admission, and predicted procedure time, both in-hospital and 30-day mortality reached statistical significance ($p = .033$ and $p = .002$, respectively). Hospital A's death rates appeared problematic for both orthopedic and general surgery, both displaying poor rank percentiles.

Overall, complications at Hospital A were generally unremarkable, typical of other hospitals with similar patients. Compared with other hospitals, Hospital A displayed a lower 30-day orthopedic complication rate (55.5 vs. 63 percent, $p < .05$), but a nonsignificant higher rate for general surgery (68 vs. 62.7 percent), so the overall complication rate was also no different between Hospital A and the remaining 216 hospitals, and no inpatient complication rate differences reached significance. Performance on failure-to-rescue (FTR) was generally poor, and it did reach statistical significance for 30-day FTR for both stratified and regression adjusted stratified results ($p = .0077$ and $p = .0029$, respectively). Hospital A ranked poorly for FTR, 97th and 93rd percentiles, for inpatient and 30-day, respectively.

Readmission rate was unremarkable, as was overall length of stay, although general surgery length of stay was elevated. However, costs suggested an interesting pattern. General surgical patient costs were very high as compared with controls, whereas orthopedic surgery costs were quite low, so that these significant results cancelled themselves out to produce a neutral picture on cost overall.

The percentage of patients utilizing the ICU was far higher than the remaining 216 hospital control rate, and the ICU stay per patient in the ICU was much longer. Finally, the surgical procedure time was 15 minutes longer ($p < .0001$) at this hospital than the controls, despite the fact that predicted surgical times between cases and controls were almost identical (137.1 vs. 136.1 minutes).

The measured patient characteristics at Hospital A and the remaining 216 hospitals are very similar; see the bottom of Table 2. Moreover, a sensitivity analysis showed that an unobserved patient characteristic would need to more than double the odds of a patient's being treated at Hospital A *and* more than double their odds of 30-day death to explain away Hospital A's significantly higher mortality rate (see Appendix for details of sensitivity analysis). Therefore, this hospital's Chief Medical Officer could not easily explain worse mortality by suggesting that their patients were somehow sicker than the other 216 hospitals.

Does the Template Reflect the Hospital's Case Mix?

In Figure 3 the distribution on the left (dashed) displays the cross-match Upsilon estimates comparing each of the 217 study hospitals to the template. Hospital A had 134 cross-matches (Upsilon = 0.45, $p = .0304$), indicating that, before matching to the template, its original patient mix differed from the template (although hospital A had abundant patients like those in the template, thus permitting a very close match to the template; see the solid line in Figure 3).

DISCUSSION

The results of our analysis of general surgery and orthopedics displayed considerable variation in outcomes, despite very uniform patient characteristics in the template for each hospital. The matched sample allows for a fair, directly standardized comparison across hospitals, since the matched sample of 300 patients is closely balanced between each hospital and the template. The resulting variation in outcomes was therefore more believable. We can say that in this study observable patient risk factors across hospitals were far more balanced than would occur in a randomized trial. A limitation of every method of standardization for observed patient characteristics is that it may fail to control some unobserved patient characteristic (Gowrisankaran and

Figure 3: Using the Cross-Match Upsilon to Examine Both the Quality of the Template Match and Its Representativeness. This figure describes the distribution of two cross-match tests in terms of the estimate of Upsilon across the 217 hospitals in the study. The dashed density plot is before matching to the template (a random sample from the hospital), while the solid density plot is after matching to the template. The solid distribution on the right represents the Upsilon statistic based on the cross-matches between the Template and the matched 300 patient sample from each individual hospital. The vertical dashed line displays the critical Upsilon where hospitals below that value display matches significantly worse than would be expected in a randomized trial. There were no such hospitals to the left of the critical region. Hospital A (large dot) displayed a typical matching quality to the Template, falling near the middle of the hospitals. The dashed curve on the left displays the Upsilon distribution between the Template and a random sample of 300 patients from each individual hospital in order to ask how “representative” the Template is of the hospital’s typical patients. Note Hospital A appears to have a distribution of patients that are different from the Template (a significant Upsilon test), but it did have enough overlapping patients to the Template so that the cross-match test between the matched patients from Hospital A and the Template were not statistically different (as seen in the solid distribution to the right)

Town 1999; Geweke, Gowrisankaran, and Town 2003). Techniques that combine matching methods with potential instruments or random experiments are being introduced (Zhang et al. 2011), although routine use for hospital auditing may be difficult to implement. However, our sensitivity analysis indicates that the elevated mortality rate in Hospital A is not easily attributed to an unobserved covariate: that covariate would need to more than double the chance of admission to Hospital A and double mortality risk. Because they are a straightforward computation, sensitivity analyses can easily be an element of routine audits.

When comparing template-matched samples across hospitals, all analyses are, in effect, severity adjusted. This may be a useful property when examining individual processes that may have contributed to the outcome. The matched sample from each hospital can be formally examined for fairness, as we did in this study, using the cross-match test (Heller et al. 2010a; Heller, Rosenbaum, and Small 2010b; Rosenbaum 2005) and Fisher's Exact Test. Furthermore, results for any hospital may also be presented with postmatch adjustments, as we demonstrated (Cochran and Rubin 1973; Rubin 1979; Silber et al. 2005). It is important to note that in almost all instances, when we observed outcomes that reached statistical significance in the stratified tests, they also reached a similar level of significance using the postmatch regression adjusted stratified analysis, indicating that most of the "adjustment" in template matching occurs through the match.

We also examined the difference between typical regression methods to achieve indirect standardization (by ranking hospitals on O/E) and the template method. We did not find high correlation between the two methods. Given the great variation in patient characteristics across the 217 hospitals when using the risk-only match and the lack of variation in patient characteristics using the template match, it would seem that variation in outcomes across hospitals would be far more credible using the template matching framework.

Finally, we should mention the potential use of the cross-match test to examine how closely a random sample of an index hospital's patients resembles the template ("representative"). The template is a fair exam—every hospital takes the same exam—but it is reasonable to ask whether this fair exam represents the typical patient mix at the hospital. To press the analogy, think of a physics major writing a genuinely terrible English essay. If Hospital A performs poorly in the template match, which is derived from the patient mix of hundreds of hospitals, then a typical patient at those hundreds of hospitals would be better served by avoiding Hospital A. That being said, it would also

be interesting to examine how that same hospital performed using a template constructed of its own patients used for constructing matched samples across other hospitals that could match this index hospital's template. We call such a match "direct-indirect" standardization and provide details of this comparison in another paper (Silber et al. 2014). A hospital that performs poorly on the external template but well (relative to other hospitals) on its own template can at least argue that they provide good care to the types of patients they generally see, just not to patients chosen for analysis by the external template. There are situations when we may want to ask, "How is the hospital doing relative to other hospitals with the patients it sees," and this question may best be answered with indirect standardization. There are other situations when we may want to ask, "How is the hospital doing relative to other hospitals on a defined set of patients of interest," and this question is best answered with direct standardization.

We believe that use of a template match will bring new opportunities to better evaluate hospital quality of care. We compare similar patients at different hospitals, a fair exam. When a hospital does poorly on the exam, it can understand its poor performance in terms of specific patients at its hospital matched to the template, and on that basis it can look for the causes of its poor performance within its institutional structure. With template matching, we believe Chief Medical Officers will be better able to examine how their hospitals are doing, and why they are achieving the results they observe, with less concern about differences in patients across hospitals.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: We thank Traci Frank and Alexander Hill for their assistance on this project. This research was funded by grant R01-HS018338 from the Agency for Healthcare Research and Quality and grant NSF SBS-10038744 from the National Science Foundation.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Berwick, D., and D. L. Wald. 1990. "Hospital Leaders' Opinions of the HCFA Mortality Data." *Journal of the American Medical Association* 263 (2): 247–9.

- Cochran, W. G., and D. B. Rubin. 1973. "Controlling Bias in Observational Studies. A Review." *Sankhya* 35 (4): 417–46.
- Fleiss, J. L., B. Levin, and M. C. Paik. 2003. "Chapter 19. The Standardization of Rates." In *Statistical Methods for Rates and Proportions*, pp. 627–647. New York: John Wiley & Sons.
- Geweke, J., G. Gowrisankaran, and R. J. Town. 2003. "Basyesian Inference for Hospital Quality in a Selection Model." *Econometrica* 71 (4): 1215–38.
- Gowrisankaran, G., and R. J. Town. 1999. "Estimating the Quality of Care in Hospitals Using Instrumental Variables." *Journal of Health Economics* 18 (6): 747–67.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. "Chapter 6. Linear Models: Robust Estimation." In *Robust Statistics. The Approach Based on Influence Functions*, pp. 315–28. New York: John Wiley & Sons.
- Hansen, B. B. 2004. "Full Matching in an Observational Study of Coaching for the Sat." *Journal of the American Statistical Association* 99 (467): 609–18.
- . 2007. "Optmatch: Flexible, Optimal Matching for Observational Studies." *R News* 7 (2): 18–24.
- Hansen, B. B., and S. O. Klopfer. 2006. "Optimal Full Matching and Related Designs Via Network Flows." *Journal of Computational and Graphical Statistics* 15 (3): 609–27.
- Heller, R., P. R. Rosenbaum, and D. Small. 2010b. "Using the Cross-Match Test to Appraise Covariate Balance in Matched Pairs." *The American Statistician* 64 (4): 299–309.
- Heller, R., S. T. Jensen, P. R. Rosenbaum, and D. S. Small. 2010a. "Sensitivity Analysis for the Cross-Match Test, with Applications in Genomics." *Journal of the American Statistical Association* 105 (491): 1005–13.
- Huber, P. J. 1981. *Robust Statistics*. Hoboken, NJ: John Wiley & Sons.
- Iezzoni, L. I. 2012. *Risk Adjustment for Measuring Health Care Outcomes*, 4th edition. Chicago, IL: Health Administration Press.
- Kitagawa, E. M. 1955. "Components of a Difference between Two Rates." *Journal of the American Statistical Association* 50 (272): 1168–94.
- Kruskal, W., and W. A. Wallis. 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260): 583–621.
- Lehmann, E. L. 2006. "Chapter 3. Blocked Comparisons for Two Treatments. Section 3. Combining Data from Several Experiments or Blocks." In *Nonparametrics: Statistical Methods Based on Ranks*, pp. 132–41. New York: Springer.
- Miettinen, O. S. 1969. "Individual Matching with Multiple Controls in the Case of All-or-None Responses." *Biometrics* 25 (2): 339–55.
- R Development Core Team. 2012. "R: A Language and Environment for Statistical Computing" [accessed on May 14, 2013]. Available at <http://www.R-project.org>
- Rosenbaum, P. R. 1988. "Sensitivity Analysis for Matching with Multiple Controls." *Biometrika* 75 (3): 577–81.
- . 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society B* 53 (3): 597–610.
- . 2002. *Observational Studies*, 2nd edition. New York: Springer-Verlag.

- . 2005. "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency." *Journal of the Royal Statistical Society B* 67 (Pt 4): 515–30.
- . 2010a. "Chapter 10: Fine Balance." In *Design of Observational Studies*, pp. 197–206. New York: Springer.
- . 2010b. *Design of Observational Studies*. New York: Springer.
- Rosenbaum, P. R., R. N. Ross, and J. H. Silber. 2007. "Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association* 102 (477): 75–83.
- Rosenbaum, P. R., and J. H. Silber. 2009. "Amplification of Sensitivity Analysis in Matched Observational Studies." *Journal of the American Statistical Association* 104 (488): 1398–405.
- Rubin, D. B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74 (366): 318–28.
- . 1980. "Bias Reduction Using Mahalanobis Metric Matching." *Biometrics* 36 (2): 293–8.
- . 2007. "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.
- . 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2 (3): 808–40.
- Silber, J. H., P. R. Rosenbaum, and R. N. Ross. 1995. "Comparing the Contributions of Groups of Predictors: Which Outcomes Vary with Hospital Rather Than Patient Characteristics?" *Journal of the American Statistical Association* 90 (429): 7–18.
- Silber, J. H., S. V. Williams, H. Krakauer, and J. S. Schwartz. 1992. "Hospital and Patient Characteristics Associated with Death after Surgery: A Study of Adverse Occurrence and Failure to Rescue." *Medical Care* 30 (7): 615–29.
- Silber, J. H., P. R. Rosenbaum, M. E. Trudeau, W. Chen, X. Zhang, S. A. Lorch, R. R. Kelz, R. E. Mosher, and O. Even-Shoshan. 2005. "Preoperative Antibiotics and Mortality in the Elderly." *Annals of Surgery* 242 (1): 107–14.
- Silber, J. H., P. S. Romano, A. K. Rosen, Y. Wang, R. N. Ross, O. Even-Shoshan, and K. Volpp. 2007a. "Failure-to-Rescue: Comparing Definitions to Measure Quality of Care." *Medical Care* 45 (10): 918–25.
- Silber, J. H., P. R. Rosenbaum, X. Zhang, and O. Even-Shoshan. 2007b. "Estimating Anesthesia and Surgical Procedure Times from Medicare Anesthesia Claims." *Anesthesiology* 106 (2): 346–55.
- Silber, J. H., P. R. Rosenbaum, D. Polsky, R. N. Ross, O. Even-Shoshan, J. S. Schwartz, K. A. Armstrong, and T. C. Randall. 2007c. "Does Ovarian Cancer Treatment and Survival Differ by the Specialty Providing Chemotherapy?" *Journal of Clinical Oncology* 25 (10): 1169–75.
- Silber, J. H., P. R. Rosenbaum, X. Zhang, and O. Even-Shoshan. 2007d. "Influence of Patient and Hospital Characteristics on Anesthesia Time in Medicare

- Patients Undergoing General and Orthopedics Surgery.” *Anesthesiology* 106 (2): 356–64.
- Silber, J. H., P. R. Rosenbaum, O. Even-Shoshan, L. Mi, F. Kyle, Y. Teng, D. W. Bratzler, and L. A. Fleisher. 2011. “Estimating Anesthesia Time Using the Medicare Claim: A Validation Study.” *Anesthesiology* 115 (2): 322–33.
- Silber, J. H., P. R. Rosenbaum, R. R. Kelz, C. E. Reinke, M. D. Neuman, R. N. Ross, O. Even-Shoshan, G. David, P. A. Saynisch, F. A. Kyle, D. W. Bratzler, and L. A. Fleisher. 2012. “Medical and Financial Risks Associated with Surgery in the Elderly Obese.” *Annals of Surgery* 256 (1): 79–86.
- Silber, J. H., P. R. Rosenbaum, A. S. Clark, B. J. Giantonio, R. N. Ross, Y. Teng, M. Wang, B. A. Niknam, J. M. Ludwig, W. Wang, O. Even-Shoshan, and K. R. Fox. 2013a. “Characteristics Associated with Differences in Survival among Black and White Women with Breast Cancer.” *Journal of the American Medical Association* 310 (4): 389–97.
- Silber, J. H., P. R. Rosenbaum, R. N. Ross, O. Even-Shoshan, R. R. Kelz, M. D. Neuman, C. E. Reinke, J. M. Ludwig, F. A. Kyle, D. W. Bratzler, and L. A. Fleisher. 2013b. “Racial Disparities in Operative Procedure Time: The Influence of Obesity.” *Anesthesiology* 119 (1): 43–51.
- Silber, J. H., P. R. Rosenbaum, R. N. Ross, J. M. Ludwig, W. Wang, B. A. Niknam, P. A. Saynisch, O. Even-Shoshan, R. R. Kelz, and L. A. Fleisher. 2014. “Hospital-Specific Template Matching for Auditing Hospital Cost and Quality. (Companion Paper).” *Health Services Research*.
- Yang, D., D. S. Small, J. H. Silber, and P. R. Rosenbaum. 2012. “Optimal Matching with Minimal Deviation from Fine Balance in a Study of Obesity and Surgical Outcomes.” *Biometrics* 68 (2): 628–36.
- Zhang, K., D. Small, S. A. Lorch, S. Srinivas, and P. R. Rosenbaum. 2011. “Using Split Samples and Evidence Factors in an Observational Study of Neonatal Outcomes.” *Journal of the American Statistical Association* 106 (494): 511–24.
- Zubizarreta, J. R. 2012. “Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery.” *Journal of the American Statistical Association* 107 (500): 1360–71.
- Zubizarreta, J. R., M. Cerda, and P. R. Rosenbaum. 2013. “Effect of the 2010 Chilean Earthquake on Posttraumatic Stress: Reducing Sensitivity to Unmeasured Bias through Study Design.” *Epidemiology* 24 (1): 79–87.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Table S1: Expanded Version of Table 1 (Assessing Whether Patient Covariates and Outcomes Vary Significantly across Hospitals).

Table S2: Expanded Version of Table 2 (Understanding an Individual Hospital's Outcomes).

Table S3: Description of the Template Sample.

Table S4: Definitions of Index Surgical Populations and Hospitals.

Table S5: Description of Matching Algorithm.

Table S6: Description of Predicted Procedure Time Algorithm. (a) Models Created to Predict Procedure Time. (b) Estimates, in Minutes, for All ICD-9-CM Principal Procedure-Secondary Procedure Interactions.

Table S7: Description of Procedure Group-specific Models to Predict Thirty day Mortality.

Table S8: Definitions of ICD-9-CM Procedure Groups and Clusters.

Table S9: Definitions and Groupings of ICD-9-CM Secondary Procedures.

Table S10: Sensitivity Analysis.

Appendix SA1: Author Matrix.