ORIGINAL ARTICLE

Introduction to Propensity Scores

A Case Study on the Comparative Effectiveness of Laparoscopic vs Open Appendectomy

Mark R. Hemmila, MD; Nancy J. Birkmeyer, PhD; Saman Arbabi, MD, MPH; Nicholas H. Osborne, MD; Wendy L. Wahl, MD; Justin B. Dimick, MD, MPH

Objective: To demonstrate the use of propensity scores to evaluate the comparative effectiveness of laparoscopic and open appendectomy.

Design: Retrospective cohort study.

Setting: Academic and private hospitals.

Patients: All patients undergoing open or laparoscopic appendectomy (n=21 475) in the Public Use File of the American College of Surgeons National Surgical Quality Improvement Program were included in the study. We first evaluated the surgical approach (laparoscopic vs open) using multivariate logistic regression. We next generated propensity scores and compared outcomes for open and laparoscopic appendectomy in a 1:1 matched cohort. Covariates in the model for propensity scores included comorbidities, age, sex, race, and evidence of perforation.

Main Outcome Measures: Patient morbidity and mortality, rate of return to operating room, and hospital length of stay.

Results: Twenty-eight percent of patients underwent open appendectomy, and 72% had a laparoscopic approach; 33% (open) vs 14% (laparoscopic) had evidence of a ruptured appendix. In the propensity-matched cohort, there was no difference in mortality (0.3% vs 0.2%), reoperation (1.8% vs 1.5%), or incidence of major complications (5.9% vs 5.4%) between groups. Patients undergoing laparoscopic appendectomy experienced fewer wound infections (odds ratio [OR], 0.4; 95% confidence interval [CI], 0.3-0.5) and fewer episodes of sepsis (0.8; 0.6-1.0) but had a greater risk of intra-abdominal abscess (1.7; 1.3-2.2). An analysis using multivariate adjustment resulted in similar findings.

Conclusions: After accounting for patient severity, open and laparoscopic appendectomy had similar clinical outcomes. In this case study, propensity score methods and multivariate adjustment yielded nearly identical results.

Arch Surg. 2010;145(10):939-945

ENTHUSIASM FOR "COMPARATIVE effectiveness" research is at an all-time high. Comparative effectiveness research is aimed at providing information on the relative strengths and weaknesses of various medical treatments. Randomized clinical trials are widely heralded as the best method of evaluating the efficacy of medical treatments. The randomization process ensures that the 2 treatment groups are balanced for all potential patient characteristics and makes certain that inferences about the effectiveness of the treatment are not threatened by confounding variables.

See Invited Critique at end of article

However, randomized trials evaluate only the efficacy of the treatment in a narrow context and do not provide information on the effectiveness of the interventions when they are applied to a wider population. To evaluate interventions in the real world, we often must rely on observational studies. The Achilles heel of observational studies is potential confounding by differences in patient characteristics. Unlike randomized trials, where the chance assignment of treatment balances patient characteristics, there exists significant potential for selection bias in observational studies.

Studies that use propensity scores are appearing with increasing frequency in the surgical literature.[1,2] Propensity scores are a statistical technique for dealing with selection bias in observational studies.[3,4] Selection bias arises when certain types of patients are more or less likely to receive treatment owing to possible confounding by indication. For example, when selecting a surgical approach for appendicitis, physicians who suspect perforation may be more likely to perform an open (vs a laparoscopic) appendectomy. Thus, comparison of outcomes in the laparoscopic vs open group would be

Author Affiliations: Departments of Surgery, University of Michigan Medical School, Ann Arbor (Drs Hemmila, Birkmeyer, Osborne, Wahl, and Dimick) and University of Washington School of Medicine, Seattle (Dr Arbabi).

Table 1. Minor and Major Complication Groups

Complications
Minor
Superficial incisional surgical site infection
Progressive renal insufficiency
Urinary tract infection
Peripheral nerve injury
Deep venous thrombosis/thrombophlebitis
Major
Deep incisional surgical site infection
Organ/space surgical site infection
Wound disruption
Pneumonia
Unplanned intubation
Pulmonary embolism
Receiving mechanical ventilation for >48 h
Acute renal failure
Stroke/cerebral vascular accident
Coma >24 h
Cardiac arrest requiring cardiopulmonary resuscitation
Myocardial infarction
Bleeding >4 U of packed red blood cells
Sepsis
Septic shock

confounded by this difference in selection; that is, we would expect worse outcomes in the open group, even if there was no true difference between the approaches.

Given the growing use of this technique in the literature, it is important for surgeons to be familiar with propensity score analysis. With propensity scores, patient and provider characteristics are used to calculate the probability that a patient will receive the intervention.[5,6] These scores are then added to multivariate models to risk adjust the analyses. Alternatively, propensity scores can be used to create matched patient cohorts.[6-8] Both approaches aim to adjust for or balance patient characteristics, thus minimizing confounding due to selection bias.

In the present study, we use the comparison of laparoscopic vs open appendectomy as a case study to introduce propensity scores. Study of the surgical treatment of this disease is ideal for the use of propensity scores because the choice of technique is often confounded by observable factors, such as severity of illness and the presence of appendiceal perforation. Thus, an unadjusted comparison of the 2 techniques will yield inaccurate results. Indeed, the laparoscopic approach has had much more favorable results in recent large observational studies[9,10] than in randomized clinical trials. To perform this case study, we used data from the Public Use File of the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP). We present an evaluation of laparoscopic vs open appendectomy for acute appendicitis.

METHODS

PATIENT DATA

This study was performed using 2005 to 2007 data from the Public Use File of the ACS-NSQIP. The study cohort consisted of patients with the postoperative diagnosis of acute appendicitis who received either an open or a laparoscopic procedure to resect the appendix. Specifically, patients were included who had a *Current Procedural Terminology (CPT)* code of 44950 (appendectomy), 44960 (appendectomy for ruptured appendix with abscess or generalized peritonitis), 44070 (laparoscopy, surgical, appendectomy), or 44979 (unlisted laparoscopy procedure, appendix) recorded as the principal operative procedure and a postoperative *International Classification of Diseases, Ninth Revision, Clinical Modification,* code of 540 (acute appendicitis), 540.0 (acute appendicitis with generalized peritonitis), 540.1 (acute appendicitis with peritoneal abscess), 540.9 (acute appendicitis without mention of peritonitis), 541 (appendicitis, unqualified), or 542 (other appendicitis). Patients who underwent exploratory laparotomy for acute appendicitis were excluded. In this "aggregate cohort," 2 groups were formed. Patients in the open appendectomy group underwent an operative procedure with a *CPT* code of 44950 or 44960. The laparoscopic group consisted of patients who had *CPT* code 44070 or 44979 recorded as their principal operative procedure. Evidence of appendiceal perforation or rupture was defined by a *CPT* code of 44960 or an *International Classification of Diseases, Ninth Revision, Clinical Modification,* code of 540.0 to 540.1.

ANALYSIS OF AGGREGATE COHORT

Data were compared using univariate and multivariate statistical measures. Continuous variables were analyzed using an unpaired 2-tailed *t* test for data with a normal distribution. Continuous data exhibiting a skewed distribution, such as length of stay (LOS), were analyzed using the Wilcoxon rank sum test. Discrete variables were compared using a χ^2 analysis. Multivariate analysis was performed using multiple linear regression or logistic regression and adjustment for significant covariates to generate risk-adjusted outcomes. All covariates with a $P<.20$ based on univariate analysis were entered into the forward stepwise regression model. The significance-level criterion for entry into the regression model was 0.1 and for removal was 0.2. Before multivariate analysis, continuous right-skewed data were natural log transformed, the regression analysis was then conducted, and the coefficient from the regression model was exponentiated to determine the proportionate increase in LOS associated with the selected treatment group. A previously created complications classification system divided complications into 2 groups, minor and major.[11,12] A list of the complications classification system is in **Table 1**. Major complications were those considered significant enough to result in increases to the LOS or a need for substantial additional treatment interventions. All statistical analysis was performed using a software program (STATA SE 9.2; StataCorp LP, College Station, Texas). Results are presented as mean (SD) unless otherwise noted. Statistical significance was defined as $P<.05$.

CREATION AND ANALYSIS OF THE PROPENSITY-MATCHED COHORT

Propensity scores were generated for surgical technique using nonparsimonious logistic regression and adjusting for important known baseline covariates, including evidence of perforation or rupture. All the covariates were entered into a logistic regression analysis, and a maximum-likelihood probit model was fitted based on these covariates as predictors of surgical technique. The probit coefficients for these predictors of surgical technique were used to calculate a propensity score of 0 to 1 for each patient. Based on the calculated propensity scores, 2 evenly matched groups were formed regarding surgical tech-

Table 2. Patient Characteristics

Patient Characteristic	Aggregate Cohort			Propensity-Matched Cohort		
	Open Appendectomy	Laparoscopic Appendectomy	*P* Value	Open Appendectomy	Laparoscopic Appendectomy	*P* Value
No. (%)	6030 (28)	15 445 (72)		5666 (50)	5666 (50)	
Age, mean (SD), y	41 (17)	38 (16)	<.001	40.1 (16.8)	41.4 (17.2)	<.001
Female, No. (%)	2551 (42)	7458 (48)	<.001	2425 (43)	2495 (44)	.20
Nonwhite race, No. (%)	2306 (38)	5451 (35)	<.001	2123 (37)	2132 (38)	.90
ASA class, No. (%)						
1-2	5139 (85)	14 005 (91)	<.001	4944 (87)	4854 (86)	.02
3-5	891 (15)	1440 (9)		722 (13)	812 (14)	
Emergency surgery, No. (%)	4938 (82)	11 884 (77)	<.001	4597 (81)	4571 (81)	.50
Wound class, No. (%)						
Clean-contaminated	2198 (36.8)	6036 (39.7)	<.001	2207 (39.0)	2163 (38.2)	.40
Contaminated	1742 (29.2)	6467 (42.5)		1746 (30.8)	1706 (30.1)	
Dirty/infected	2031 (34.0)	2707 (17.8)		1711 (30.2)	1794 (31.7)	
Evidence of rupture (*CPT* code 44960 or *ICD-9-CM* codes 540.0 and 540.1), No. (%)	1976 (33)	2136 (13.8)	<.001	1628 (29)	1733 (31)	.03
Selected comorbid risk factors, No. (%)						
No diabetes	5703 (94.6)	14 833 (96.0)	<.001	5399 (95.3)	5345 (94.3)	.03
Current smoker	1319 (21.9)	3426 (22.2)	.60	1247 (22.0)	1347 (23.8)	.03
Ethanol use	187 (3.1)	302 (2.0)	<.001	158 (2.8)	182 (3.2)	.20
No dyspnea	5883 (97.6)	15 216 (98.5)	<.001	5541 (97.8)	5540 (97.8)	.90
DNR	30 (0.5)	40 (0.3)	.006	25 (0.4)	24 (0.4)	.90
Independent functional status	5851 (97.0)	15 194 (98.4)	<.001	5535 (97.7)	5519 (97.4)	.40
History of severe COPD	84 (1.4)	125 (0.8)	<.001	67 (1.2)	73 (1.3)	.60
Ascites within 30 d	124 (2.1)	282 (1.8)	.30	105 (1.9)	122 (2.2)	.30
History of MI	23 (0.4)	20 (0.1)	<.001	16 (0.3)	16 (0.3)	>.99
Hypertension	1187 (19.7)	2222 (14.4)	<.001	1012 (17.9)	1189 (21.0)	<.001
Acute renal failure	17 (0.3)	13 (0.08)	<.001	8 (0.1)	8 (0.1)	>.99
Currently undergoing dialysis	30 (0.5)	25 (0.2)	<.001	15 (0.3)	14 (0.3)	.90
Sepsis						
SIRS	2188 (36.3)	5178 (33.6)	<.001	2037 (36.0)	2078 (36.7)	.80
Sepsis	187 (3.1)	205 (1.3)		138 (2.4)	145 (2.6)	
Septic shock	31 (0.5)	21 (0.1)		14 (0.3)	14 (0.3)	
Pregnancy	76 (1.3)	144 (0.9)	.007	14 (0.3)	13 (0.2)	.80

Abbreviations: ASA, American Society of Anesthesiologists; COPD, chronic obstructive pulmonary disease; *CPT, Current Procedural Terminology*; DNR, do not resuscitate; *ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification*; MI, myocardial infarction; SIRS, systemic inflammatory response syndrome.

nique using a matching algorithm with the common caliper set at 0.005. Caliper is the maximum distance or difference that is acceptable for a propensity score match. The matching approach technique of using propensity scores, as opposed to stratification or regression adjustment, was chosen because it is the closest approximate to a randomized clinical trial and provides the greatest balance between treated and untreated cases.[7] This data set is referred to as the "propensity-matched cohort." The matched cohort was evaluated for balance between the 2 surgical technique groups regarding each of the potential confounding factors. Differences in outcomes such as mortality, hospital LOS, and complications were explored using univariate tests. Additional risk adjustment was performed using multivariate analysis to adjust for covariates that remained unbalanced between the 2 groups after propensity score matching.

DISCLOSURE

The ACS-NSQIP and the hospitals participating in the ACS-NSQIP are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors. Approval for this study was obtained from the University of Michigan Health System institutional review board.

RESULTS

PATIENT DEMOGRAPHICS

A total of 21 475 patients underwent appendectomy during the study, with 28% of these patients undergoing open appendectomy and 72% a laparoscopic approach. The characteristics of patients in the laparoscopic and open groups showed many important differences (**Table 2**). Notably, there were some key differences between the laparoscopic and open groups that could act as confounding variables. For example, patients in the open appendectomy group were more likely than those in the laparoscopic group to have evidence of a ruptured appendix (33% vs 14%). In summary, of the 41 ACS-NSQIP preoperative risk factors, 28 were found to have differences present between the 2 groups on univariate analysis (not all data shown).

UNADJUSTED EVALUATION OF OUTCOMES

In unadjusted analysis, there were large differences in outcomes between the laparoscopic and open groups (**Table 3**). All complications for which a difference was found favored

Table 3. Unadjusted Outcomes (Univariate Analysis)

Morbidity	Aggregate Cohort, No. (%)			Propensity-Matched Cohort, No. (%)		
	Open Appendectomy	Laparoscopic Appendectomy	*P* Value	Open Appendectomy	Laparoscopic Appendectomy	*P* Value
Superficial incisional SSI	266 (4.4)	207 (1.3)	<.001	243 (4.3)	99 (1.8)	<.001
Deep incisional SSI	69 (1.1)	40 (0.3)	<.001	57 (1.0)	20 (0.4)	<.001
Organ/space SSI	113 (1.9)	292 (1.9)	.90	103 (1.8)	172 (3.0)	<.001
Wound disruption	33 (0.6)	12 (0.08)	<.001	26 (0.5)	7 (0.1)	.001
Pneumonia	43 (0.7)	42 (0.3)	<.001	30 (0.5)	29 (0.5)	.90
Unplanned intubation	38 (0.6)	29 (0.2)	<.001	24 (0.4)	20 (0.4)	.50
Pulmonary embolism	6 (0.1)	15 (0.1)	.90	6 (0.1)	9 (0.2)	.40
Receiving mechanical ventilation for >48 h	45 (0.8)	21 (0.1)	<.001	24 (0.4)	15 (0.3)	.10
Progressive renal insufficiency	9 (0.2)	7 (0.05)	.01	8 (0.1)	6 (0.1)	.60
Acute renal failure	8 (0.1)	10 (0.06)	.10	3 (0.05)	10 (0.2)	.052
Urinary tract infection	34 (0.6)	64 (0.4)	.10	28 (0.5)	33 (0.6)	.50
Stroke/CVA	3 (0.05)	2 (0.01)	.10	3 (0.05)	1 (0.02)	.30
Coma >24 h	1 (0.02)	2 (0.01)	.80	1 (0.02)	1 (0.02)	>.99
Peripheral nerve injury	1 (0.02)	2 (0.01)	.80	1 (0.02)	2 (0.04)	.60
Cardiac arrest requiring CPR	11 (0.2)	6 (0.04)	.001	9 (0.2)	5 (0.1)	.30
Myocardial infarction	7 (0.1)	4 (0.03)	.009	5 (0.1)	2 (0.04)	.20
Bleeding/transfusions	1 (0.02)	5 (0.03)	.50	1 (0.02)	0	.30
DVT/thrombophlebitis	18 (0.3)	22 (0.1)	.02	12 (0.2)	11 (0.2)	.80
Sepsis	156 (2.6)	175 (1.1)	<.001	135 (2.4)	103 (1.8)	.04
Septic shock	35 (0.6)	28 (0.2)	<.001	24 (0.4)	21 (0.4)	.70

Abbreviations: CPR, cardiopulmonary resuscitation; CVA, cerebrovascular accident; DVT, deep venous thrombosis; SSI, surgical site infection.

Table 4. Comparison of Unadjusted and Adjusted Outcomes for Laparoscopic vs Open Appendectomy[a]

Outcome	Aggregate Cohort		Propensity-Matched Cohort		Aggregate Cohort				Propensity-Matched Cohort, Unadjusted	
					Unadjusted		Adjusted			
	Open Appendectomy (n=6030)	Laparoscopic Appendectomy (n=15 445)	Open Appendectomy (n=5666)	Laparoscopic Appendectomy (n=5666)	OR (95% CI)[b]	*P* Value	OR (95% CI)[b]	*P* Value	OR (95% CI)[b]	*P* Value
Mortality	26 (0.4)	17 (0.1)	14 (0.3)	13 (0.2)	0.3 (0.1-0.5)	<.001	1.05 (0.5-2.2)	.9	0.9 (0.4-2.0)	.8
Length of stay, median (IQR), d	2 (1-4)	1 (1-2)	2 (1-4)	1 (1-3)	–0.8[c]	<.001	–0.5[c]	<.001	–0.4[c]	<.001
Return to the operating room	116 (1.9)	169 (1.1)	99 (1.8)	87 (1.5)	0.6 (0.4-0.7)	<.001	0.8 (0.6-1.1)	.1	0.9 (0.7-1.2)	.4
Incisional SSI	330 (5.5)	246 (1.6)	295 (5.2)	118 (2.1)	0.3 (0.2-0.3)	<.001	0.3 (0.3-0.4)	<.001	0.4 (0.3-0.5)	<.001
Organ/space SSI	113 (1.9)	292 (1.9)	103 (1.8)	172 (3.0)	1.0 (0.8-1.3)	.9	1.9 (1.5-2.3)	<.001	1.7 (1.3-2.2)	<.001
Sepsis or septic shock	188 (3.1)	200 (1.3)	157 (2.8)	121 (2.1)	0.4 (0.3-0.5)	<.001	0.8 (0.6-0.94)	.01	0.8 (0.6-1.0)	.03
Any complication	656 (10.9)	765 (5.0)	563 (9.9)	428 (7.6)	0.4 (0.4-0.5)	<.001	0.6 (0.6-0.7)	<.001	0.7 (0.6-0.8)	<.001
Major complication	403 (6.7)	505 (3.3)	334 (5.9)	303 (5.4)	0.5 (0.4-0.5)	<.001	0.9 (0.8-0.99)	.047	0.9 (0.8-1.1)	.2
Minor complication	323 (5.4)	297 (1.9)	287 (5.1)	149 (2.6)	0.3 (0.3-0.4)	<.001	0.4 (0.4-0.5)	<.001	0.5 (0.4-0.6)	<.001

Abbreviations: CI, confidence interval; IQR, interquartile range; OR, odds ratio; SSI, surgical site infection.
[a] Data are given as number (percentage) unless otherwise indicated.
[b] All ORs are for laparoscopic vs open appendectomy.
[c] Attributable change in length of stay for laparoscopic appendectomy.

the laparoscopic technique. In the aggregate cohort, the rate of mortality, LOS, return to the operating room, incisional surgical site infection (SSI), sepsis with or without septic shock, and minor and major complications were significantly higher on univariate analysis for patients who underwent open surgery (**Table 4**).

MULTIVARIATE ADJUSTMENT

After adjustment for confounding variables using multivariate logistic regression, the differences in outcomes were entirely changed (Table 4). Operative mortality and return to the operating room were no longer different between the laparoscopic and open approaches. Patients in the laparoscopic group had a reduction in their hospital LOS and rates of incisional SSI, sepsis, and major and minor complications compared with the open operation group. However, patients who underwent laparoscopic appendectomy experienced a significantly higher rate of organ/space SSI or intra-abdominal abscess (odds ratio [OR], 1.9; 95% confidence interval [CI], 1.5-2.3). The rate of organ/space SSI was substantially higher in the mul-

tivariate-adjusted analysis compared with the univariate result for the aggregate cohort (OR, 1.9 vs 1.0) (**Figure**, B).

PROPENSITY-MATCHED COHORT

In the propensity-matched cohort, there were 5666 patients in each of the laparoscopic and open groups (Table 2). Most of the differences in patient characteristics were no longer present. There was still a detectable difference in the rate of appendiceal perforation; however, in practicality, these rates were nearly identical in the open and laparoscopic groups (29.0% vs 31.0%) compared with the large difference (33.0% vs 14.0%) in the aggregate cohort. The previously seen differences in preoperative risk factors declined from 28 to just 3 of 41 comorbid conditions.

In the propensity-matched cohort, there were fewer differences in univariate unadjusted outcomes between the 2 groups (Table 3). There were no differences in the rate of mortality, return to the operating room, or major complications. Patients who underwent open appendectomy had higher rates of incisional SSI, wound disruption, and sepsis. However, the incidence of organ/space SSI was much higher in the laparoscopic approach compared with the open approach in the propensity-matched cohort (OR, 1.7; 95% CI, 1.3-2.2). The Figure illustrates selected outcomes between the aggregate cohort without and with adjusted analysis compared with the propensity-matched cohort. The Figure also includes meta-analysis data from a 2004 Cochrane review of randomized trials for appendectomy.[13] The meta-analysis results for incisional and organ/space SSI were similar to the present results for the aggregate cohort with multivariate adjustment and the propensity-matched cohort analysis.

To further explore the differences between these 2 surgical approaches, we performed a risk-adjusted analysis of the propensity cohort. Variables with differences on univariate testing were entered into the multivariate model. Risk adjustment did not change any of the previously positive results found on univariate analysis (data not shown). Patients who underwent an open operation had an increase in their median hospital LOS by a half day attributable to the surgical technique.

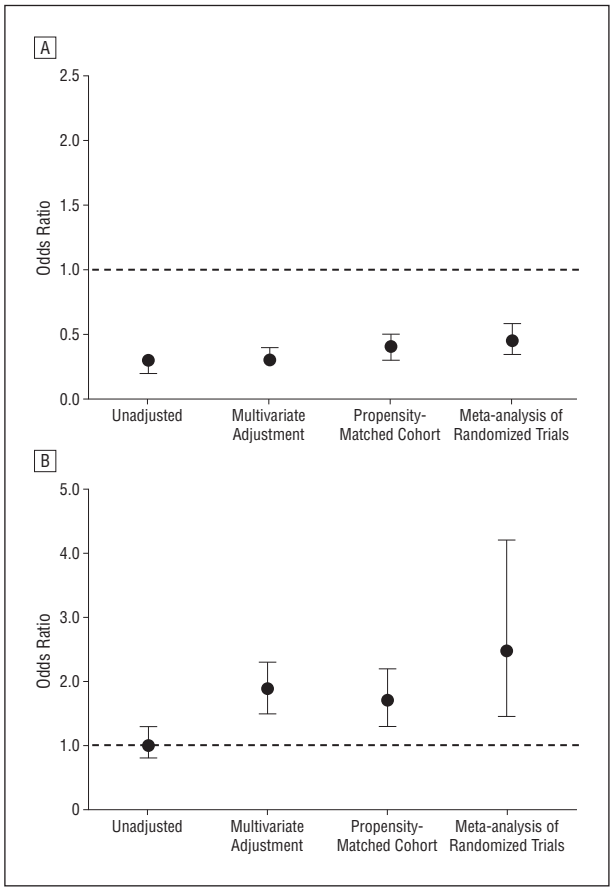

Figure. Comparison of outcomes for incisional (A) and organ/space (B) surgical site infections: aggregate cohort unadjusted, aggregate cohort risk-adjusted, propensity-matched cohort unadjusted, and meta-analysis outcomes. Error bars represent 95% confidence intervals. Meta-analysis data from Sauerland et al.[13]

COMMENT

Using appendicitis as a case study, we present an introduction to the use of propensity scores to compare 2 surgical treatments. We chose to evaluate the operative treatment for appendicitis for several reasons. First, this is a clinical scenario familiar to most surgeons, which makes it an ideal case study to introduce a new statistical technique, such as propensity scores. Second, the 2 approaches to operative treatment (laparoscopic vs open) are often applied to different groups of patients. For example, the open approach is more often applied in patients with evidence of appendiceal perforation or septicemia. Because these factors are also associated with outcomes, they can potentially act as confounding variables in the comparison of the 2 approaches. An unadjusted comparison of laparoscopic vs open appendectomy would, therefore, yield misleading results. Indeed, in the present analysis, we found differences in most outcome variables in the unadjusted comparisons.

There are several approaches to dealing with potentially confounding variables. Multivariate regression is the technique most often used to adjust for the presence of confounding variables. In this study, there were dramatic changes in the results when we applied multivariate adjustment. Many of the differences in outcomes between laparoscopic and open appendectomy were absent after this adjustment. The results of the multivariate adjustment are identical to the results of a meta-analysis of randomized trials.[13] Specifically, open appendectomy had more wound infections, but laparoscopy had more intra-abdominal abscesses. For those using observational studies to evaluate the comparative effectiveness of treatments, these results are encouraging.

We next evaluated the use of propensity scores to adjust for these confounding variables. To use this technique, a propensity score is first assigned to each patient. This score is the likelihood that the patient receives the treatment based on all observed characteristics. For this study, the score was calculated by performing a logistic regression model in which laparos-

copy was the dependent variable and all other patient characteristics were included as independent variables. The predicted probability of laparoscopy—the propensity score—can then be estimated for each patient from this model.[14] The propensity score can then be used as an additional covariate in the multivariate adjustment or can be used to create matched pairs of patients. For this study, we created a matched cohort, which is meant to simulate a randomized trial. These cohorts are well matched on all observed variables. However, there is one key difference between propensity scores and a randomized trial. In randomized trials, patients are also matched on unobserved variables. If any important unmeasured confounding variables are not captured in the data (ie, they are unobserved), the propensity score will yield a biased estimate of the treatment effect. In the present study, the propensity score analysis yielded results similar to those of a meta-analysis[13] of randomized clinical trials, implying that the important confounding variables are present in this data source, which is not surprising given the amount of detailed data collected in the ACS-NSQIP data set.

Creating case matching based on propensity scores allows for balancing of measured variables between treated and untreated patients and elimination of bias. Greater balance is typically achieved after matching directly on the propensity score rather than stratifying on quintiles of the propensity score.[7] Different methods exist for choosing which covariates to include in a propensity score model: inclusion of only true confounders, inclusion of all variables associated with the outcome, inclusion of all measured variables, and inclusion of only variables associated with treatment selection. Inclusion of only true confounders can result in up to 24% more matched pairs compared with models that include all variables or potentially weak confounders.[7] The all-variables model resulted in the nonmatching of 364 of 6030 open appendectomy cases (6.0%) with an equivalent laparoscopic case. Therefore, we did not pursue a more parsimonious propensity score model.

One key finding of this study was that the propensity score analysis was nearly identical to the analysis using multivariate adjustment. Those who advocate the use of propensity scores argue that they are superior to multivariate adjustment at addressing confounding. However, there is little evidence that propensity scores are actually better in this regard. Similar to the present study, many analyses show little difference between multivariate adjustment and propensity score adjustment. For example, Stukel and colleagues[8] compared different approaches for dealing with confounding in observational studies (multivariate adjustment and propensity scores). They evaluated the impact of cardiac catheterization on long-term acute myocardial infarction mortality and found that multivariate adjustment and propensity scores yielded similar findings, with a moderate reduction in mortality with cardiac catheterization.

However, there is one scenario in which propensity scores are always useful: when the treatment is common but the outcome of interest is rare.[15,16] When studying rare outcomes, it is sometimes not possible to use multivariate adjustment. To construct a "stable" regression model, there must be at least 10 events in the study population for each independent variable (ie, potential confounder). For example, in the present study, unplanned reintubation occurs in 0.3% of the study population (67 total events). In the unadjusted analysis, it seems that open appendectomy is associated with significantly higher rates of unplanned intubation (0.6% vs 0.2%, $P < .001$). Because there are so few events and so many potential covariates to include in the multivariate adjustment, it is impossible to create a stable logistic regression model for this outcome.

In contrast, propensity scores can still be used to perform a risk-adjusted comparison for this situation. The logistic regression model used to create propensity scores included the surgical approach (laparoscopy vs open) as the outcome, which provides more than 6000 events. The propensity score can then be used to create matched cohorts, and rates of unplanned intubation can then be compared. In contradistinction to the unadjusted results, the rates of unplanned intubation were identical in the propensity-matched cohorts (0.4% vs 0.4%, $P = .50$), highlighting the importance of this approach for dealing with confounding with rare outcome variables.

This study provides an introduction to propensity score analysis using the operative treatment of appendicitis as a case study. We found that a multivariate adjustment provided the same results as did the analysis from a propensity-matched cohort. Specifically, we found that the laparoscopic approach has a higher rate of intra-abdominal abscess and that the open approach has a higher rate of incisional SSI. These results are consistent with those of a published meta-analysis[13] of randomized clinical trials. This finding suggests that observational studies could be a valid study design for comparative effectiveness research, especially when randomized clinical trials are not feasible or when the goal is to understand the real-world impact of a treatment. Surgeons should use this case study to further understand the use of propensity scores for risk adjustment or cohort matching as increasing focus is placed on observational studies in the context of comparative effectiveness research.

Accepted for Publication: June 14, 2010.
Correspondence: Mark R. Hemmila, MD, Department of Surgery, University of Michigan Medical School, 1B407 University Hospital, 1500 E Medical Center Dr, SPC 5033, Ann Arbor, MI 48109-5033 (mhemmila@umich.edu).
Author Contributions: *Study concept and design:* Hemmila, Arbabi, and Dimick. *Acquisition of data:* Hemmila and Dimick. *Analysis and interpretation of data:* Hemmila, Birkmeyer, Osborne, Wahl, and Dimick. *Drafting of the manuscript:* Hemmila, Birkmeyer, Osborne, and Dimick. *Critical revision of the manuscript for important intellectual content:* Hemmila, Birkmeyer, Arbabi, Osborne, Wahl, and Dimick. *Statistical analysis:* Hemmila, Birkmeyer, and Dimick. *Study supervision:* Hemmila.
Financial Disclosure: None reported.
Funding/Support: This study was supported by grant K08-GM078610 from the National Institutes of Health with joint support from the American College of Surgeons and the American Association for the Surgery of Trauma (Dr Hemmila).

Previous Presentation: This paper was presented at the American College of Surgeons 95th Clinical Congress; October 12, 2009; Chicago, Illinois; and is published after peer review and revision.

REFERENCES

1. MacKenzie EJ, Rivara FP, Jurkovich GJ, et al. A national evaluation of the effect of trauma-center care on mortality. *N Engl J Med*. 2006;354(4):366-378.
2. Haas B, Jurkovich GJ, Wang J, Rivara FP, Mackenzie EJ, Nathens AB. Survival advantage in trauma centers: expeditious intervention or experience? *J Am Coll Surg*. 2009;208(1):28-36.
3. D'Agostino RB Jr. Propensity scores in cardiovascular research. *Circulation*. 2007; 115(17):2340-2343.
4. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8, pt 2):757-763.
5. Kao LS, Lally KP, Thomas EJ, Tyson JE. Improving quality improvement: a methodologic framework for evaluating effectiveness of surgical quality improvement. *J Am Coll Surg*. 2009;208(4):621-626.
6. Adamina M, Guller U, Weber WP, Oertli D. Propensity scores and the surgeon. *Br J Surg*. 2006;93(4):389-394.
7. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26(4):734-753.
8. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. 2007;297(3):278-285.
9. Guller U, Hervey S, Purves H, et al. Laparoscopic versus open appendectomy: outcomes comparison based on a large administrative database. *Ann Surg*. 2004; 239(1):43-52.
10. Faiz O, Clark J, Brown T, et al. Traditional and laparoscopic appendectomy in adults: outcomes in English NHS hospitals between 1996 and 2006. *Ann Surg*. 2008;248(5):800-806.
11. Dimick JB, Chen SL, Taheri PA, Henderson WG, Khuri SF, Campbell DA Jr. Hospital costs associated with surgical complications: a report from the private-sector National Surgical Quality Improvement Program. *J Am Coll Surg*. 2004; 199(4):531-537.
12. Hemmila MR, Jakubus JL, Maggio PM, et al. Real money: complications and hospital costs in trauma patients. *Surgery*. 2008;144(2):307-316.
13. Sauerland S, Lefering R, Neugebauer EA. Laparoscopic versus open surgery for suspected appendicitis. *Cochrane Database Syst Rev*. 2004;4(4):CD001546.
14. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17(19): 2265-2281.
15. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med*. 2002;137(8):693-695.
16. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158(3):280-287.

INVITED CRITIQUE

Are Surgeons Ready to Embrace a Paradigm Shift in Surgical Comparative Effectiveness Research?

Congratulations to Dr Hemmila and colleagues on another strong article in the arena of surgical outcomes research using novel, cutting-edge, advanced biostatistical methods. The authors are among a new breed of surgical scientists who are changing the face of surgical investigation and resident training. Cutting-edge surgical research has historically been based in basic science laboratories using bench assays, cells, and animal surgery as the standard fare. Only more recently has the world of surgical outcomes research begun to flourish. It would have been an absurd request for me to spend my research time during residency doing "outcomes research" and taking epidemiology or biostatistics classes 15 years ago. However, times are changing, and surgical educators see the benefit in training residents (and faculty) for the research they will eventually perform. I am not the only surgical faculty member who has not performed a "laboratory experiment" or written a basic science article since my research years during residency. We do clinical research; we should follow the lead of these authors and invest the time and energy needed to learn the important tools of the trade.

This project begins with an important clinical question: "Which is better: open or laparoscopic appendectomy?" It then systematically shows the reader the steps taken to answer this question using a variety of biostatistical techniques. Along the way, propensity scoring is explained in enough detail to understand the basic concepts, but not in enough depth for casual readers to perform their own similar analyses. In the typical learning approach to surgery, we have now "seen one" but I doubt that many readers are ready to "do one" or "teach one" quite yet.

As with every statistical technique, if any key assumptions are violated, the results will be biased. I raise the following questions about these critical assumptions for this propensity score–based analysis. First, is there adequate overlap between the groups? What happened to data from patients who were unable to be matched? If their data were ignored, does the study truly reflect the entire population of patients undergoing appendectomy? Second, are there any remaining residual confounders? If so, propensity score matching may actually make the situation worse because "eliminating bias (or imbalance) due to one confounder may awaken and unleash bias due to dormant, unmeasured confounders."[1]

The final questions I pose are more philosophical. What should be done when different (yet equally valid)