Statistics for the Rest of Us

Comparing apples and oranges

Eugene H. Blackstone, MD

Mitral valve repair versus replacement, internal thoracic artery versus saphenous vein graft conduits for coronary bypass, effect of chronic preoperative atrial fibrillation on outcome, gastric versus colon esophageal substitutes, complete versus incomplete off-pump revascularization, surgery in high- versus low-volume centers, balloon versus surgical aortic valvotomy. These are but a sample of studies of comparative outcome whose basis was clinical experience rather than a formal clinical trial. Often, a cursory glance at patient characteristics in each group reveals important differences that lead medical and statistical reviewers and readers alike to scoff, "They're comparing 'apples and oranges!'"

What does it take to convince the skeptic that the difference in outcome attributed to difference in treatment (or patient condition) is real? The answer to this question is not academic; it can affect the way we as physicians learn to treat our patients from studies of clinical experience.

When comparison is made in the context of a properly designed, appropriate, ethical, feasible, well-analyzed, generalizable randomized trial, most of us would accept a cause-and-effect linkage between treatment and difference in outcome. In contrast, when the comparison emanates from studies of clinical experience—ubiquitous in surgical experience and reporting—cause-and-effect attribution is considered "speculative" at best.

For 3 decades, multivariable risk factor analysis has been the mainstay for identifying and quantifying treatment outcome differences adjusted for patient characteristics. However, Kirklin and Barratt-Boyes[1] recommended that these differences be treated as *associations* with outcomes, not *causes.* There is no guarantee that risk factor analysis is an effective strategy for discovery of cause-and-effect mechanisms.[2,3]

During the 1980s, federal support for complex clinical trials in heart disease was abundant. Few of us noticed important advances being made in statistical methods for valid, nonrandomized comparisons. An example of the advances was the seminal 1983 *Biometrika* article by Paul Rosenbaum at the University of Wisconsin, Madison, and Donald Rubin at the University of Chicago, "The Central Role of the Propensity Score in Observational Studies for Causal Effects."[4] In the 1990s, as the funding climate changed, interest in methods for making nonrandomized comparisons accelerated.[5-10]

Recently, these methods have been recommended by statistical reviewers for comparative clinical studies and have been adopted by some clinical research groups. The result has been the introduction into our literature of unfamiliar methods with their unfamiliar terminology. Rather than being relieved that at last apples-to-apples comparisons can be made with rigor, medical and sometimes statistical reviewers, as well as readers, have become bewildered!

From the Departments of Thoracic and Cardiovascular Surgery and Biostatistics and Epidemiology, The Cleveland Clinic Foundation, Cleveland, Ohio.

Received for publication Dec 13, 2000; accepted for publication July 31, 2001.

Address for reprints: Eugene H. Blackstone, MD, The Cleveland Clinic Foundation, 9500 Euclid Ave, Desk F25, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2002;123:8-15

0022-5223/2002 $35.00 + 0 **12/1/120329**

doi:10.1067/mtc.2002.120329

Therefore, my purpose is to (1) clarify the nature of the problem in nonrandomized comparisons that gives rise to apples-and-oranges skepticism; (2) review previous attempts to solve the problem; (3) present a method known as *balancing scores* that can achieve apples-to-apples comparisons under some nonrandomized conditions; (4) describe in nontechnical detail construction of the simplest balancing score, the *propensity score;* (5) demonstrate how the propensity score is used; and (6) discuss limitations, pitfalls, and alternatives.

Nature of the Problem

Except by chance, characteristics differ among patients constituting comparison groups of interest in nonrandomized studies. (For lack of a better term, I use the phrase *comparison group of interest* throughout the text to indicate either a treatment or procedure difference of interest or a patient characteristic difference of interest, such as whether a patient is in chronic atrial fibrillation). These differences in characteristics between groups are often large, systematic, and statistically significant. They arise from clinically motivated patient selection. (How often does the clinical inferences section of a journal article begin, "In carefully selected patients. . . ?") They arise for undocumented reasons called "treatment variance." They sometimes arise by chance. In whatever way they arise, they invalidate direct comparisons.

For example, Table 1 contrasts a few characteristics of patients referred for stress echocardiography who reported they either were or were not receiving long-term aspirin therapy. A clinically relevant question might be, "Does long-term aspirin use convey a survival benefit, and if so, for whom?" However, a glance at the table of patient characteristics makes the reader justifiably suspicious of attributing outcome difference to aspirin treatment in such obviously selected patients. "True, true, and unrelated," says one. "Apples and oranges," says another.

Comparisons based on well-designed randomized studies provide at least 6 protections not available to the clinical investigator that increase the cause-effect believability of a comparison.[11,12] (1) Entry and exclusion criteria are prescribed and identical for the groups being compared; thus, the variables used to assign treatment are known. (2) All patients have a specified chance of receiving each treatment, avoiding both obvious and nonobvious clinical selection of patients for one treatment or the other. (3) Treatments are concurrent, avoiding temporal trends. (4) Data collection is concurrent, uniform, and high quality, eliminating differences in definition or types of variables collected. (5) Unrecorded variables affecting outcome are nearly equally distributed between groups, eliminating confounding (one of the most important benefits of randomization). (6) Assumptions underlying statistical comparison tests are met.

TABLE 1. Selected patient characteristics according to long-term aspirin use in patients undergoing stress echocardiography for known or suspected coronary artery disease

Patient characteristic	ASA (n = 2455)	No ASA (n = 4072)	*P*
Men (%)	49	56	.001
Age (y, mean ± SD)	62 ± 11	56 ± 12	<.0001
Smoker (%)	10	13	.001
Resting heart rate (beats/min)	74 ± 13	78 ± 14	<.0001
Ejection fraction (%)	50 ± 9	53 ± 7	<.0001

ASA, Long-term aspirin use; *SD,* standard deviation.

None of these protections is available in making nonrandomized comparisons. So, why not mount randomized trials for every question? Without elaborating the limitations of randomized trials (but pointing out that some comparisons, such as whether or not a person goes into atrial fibrillation, cannot be randomized), let us acknowledge that it is impossible to mount a randomized trial to address every comparison.[13]

Can anything be done to increase the credibility of comparative studies based on clinical experience rather than randomized trials?

Previous Attempts to Address the Problem

Matching

A possibly familiar method for making nonrandomized comparisons is the case-control study.[14,15] The method seems logical and straightforward in concept. Patients in one treatment group (cases) are matched with one or more patients in the other treatment group (controls) according to variables such as age, sex, and ventricular function. However, case matching is rarely easy in practice. How close in age is acceptable? How close in ejection fraction? "We don't have anyone to match this patient in both age and ejection fraction!" The more variables that need to be matched, the more difficult it is to find a match in all specified characteristics! Yet, matching on only a few variables may not protect well against apples-and-oranges comparisons.[16-18] Diabolically, selection factor effects (called *bias*), which case-matching is intended to reduce, may *increase* bias if unmatched cases are simply eliminated.[19]

Multivariable Analysis

Treatment differences in outcome may instead be identified by multivariable analysis. Such analyses examine many variables simultaneously, including the comparison variable of interest. If one is fortunate, multivariable analysis will eliminate selection factors and provide an accurate assessment of the effect of the comparison variable of interest,

TABLE 2. Selected patient characteristics according to long-term aspirin use in patients undergoing stress echocardiography for known or suspected coronary artery disease

	Quintile I		Quintile II		Quintile III		Quintile IV		Quintile V	
Patient characteristic	ASA (n = 113)	No ASA (n = 1092)	ASA (n = 194)	No ASA (n = 1111)	ASA (n = 384)	No ASA (n = 922)	ASA (n = 719)	No ASA (n = 586)	ASA (n = 1045)	No ASA (n = 261)
Men (%)	22	22	57	63	74	71	78	78	88	87
Age (y)	55	49	56	55	61	61	62	64	63	65
Smoker (%)	15	13	15	15	12	11	11	13	7	9
Resting heart rate (beats/min)	84	83	79	79	76	76	76	76	71	73
Ejection fraction (%)	53	54	54	54	53	53	49	49	49	48

Patients are grouped in quintiles according to a balancing (propensity) score. *ASA,* Long-term aspirin use.

properly adjusted for patient characteristic differences. However, until now there has been no test to determine whether we have been fortunate.[2,18,20,21]

Balancing Scores to the Rescue

Apples-to-apples nonrandomized comparisons of outcome can be achieved, within certain limitations, by use of so-called *balancing scores.*[4] Balancing scores are a class of multivariable statistical methods that identify patients with similar chances of receiving one or the other treatment, permitting nonrandomized comparisons of treatment outcomes.

The developers of balancing score methods claim that the difference in outcome between patients who have a similar balancing score, but receive different treatments, provides an unbiased estimate of the effect attributable to the comparison variable of interest.[4] That is technical jargon for saying that the method can identify the apples from among the mixed fruit of clinical practice variance, transforming an apples-to-oranges outcomes comparison into an apples-to-apples comparison.[22-25]

Astonishing!

Why Is It Called a Balancing Score?

Randomly assigning patients to alternative treatments in clinical trials balances both patient characteristics (at least in the long run) and number of subjects in each treatment arm. In a nonrandomized setting, neither patient characteristics nor number of patients is balanced for each treatment. A balancing score achieves *local* balance in patient characteristics at the expense of unbalancing n.

Table 2 illustrates local balance of patient characteristics achieved by using a specific balancing score known as the *propensity score* (see below for details). The propensity score quantified each patient's probability (propensity) of being on long-term aspirin therapy. Patients were divided into 5 equal-sized groups called *quintiles,* on the basis of having similar propensity scores (use of quintiles has a statistical rationale).[4] Simply by virtue of having similar propensity scores, patients within each quintile were found to have similar characteristics (except for age in quintile I). As might be expected, patient characteristics differed importantly from one quintile to the next; for example, most of quintile I was women; most of quintile V was men. These quintiles look like 5 individual randomized trials with differing entry and exclusion criteria, which is exactly what balancing scores are intended to achieve! Thus, the propensity score balanced essentially all patient characteristics within *localized* subsets of patients.

To achieve this balance, a widely dissimilar number of patients *actually* received long-term aspirin therapy from quintile to quintile. Quintile I contained only a few patients who received long-term aspirin therapy, whereas quintile V had few *not* receiving aspirin. Thus, balance in patient characteristics was achieved by unbalancing n.

Propensity Score

The most widely used balancing score is the propensity score.[4] For each patient, it provides an estimate of the propensity toward (probability of) belonging to one group versus another (*group membership*). In this section I will describe (1) constructing a propensity model, (2) calculating a propensity score for each patient using the propensity model, and (3) using the propensity score in various ways for balancing.

Hard Hat Area: Propensity Model Construction

For a 2-group comparison, multivariable logistic regression is used to identify factors predictive of group membership.[4] In most respects, this is what cardiothoracic groups have done for years: find correlates of (risk factors for) an event. In this case, the event is *actual* membership in one or the other comparison group of interest.

I recommend initially formulating a parsimonious explanatory model that identifies the common denominators of group membership. *Parsimonious* means "simple,"

meaning a model limited to factors deemed statistically significant. *Model* means a mathematical representation or *equation*. (See the incremental risk factor concept in chapter 6 of *Cardiac Surgery.*[1])

Once this traditional modeling is completed, a further step is taken to generate the *propensity model*. The traditional model is augmented by other factors, even if not statistically significant. Thus, the propensity model is not parsimonious.[22] The goal is to balance patient characteristics by incorporating "everything" recorded that may relate to either systematic bias or simply bad luck.[17]

When taken to the extreme, forming the propensity model can cause problems, because medical data tend to have many variables that measure the same thing. The solution is to pick one variable from among a closely related cluster of variables as a representative of the cluster. For example, select one variable representing body size from among height, weight, body surface area, and body mass index.

When a propensity model is being formed, information should not be thrown away. Some biostatistical collaborators dichotomize (group) continuous variables, such as age or weight. This throws away information. Rather, the propensity model should incorporate *continuous* variables so as to produce a smooth distribution of scores necessary for good local matching.

Other construction tips are presented in the appendix.

Calculating the Propensity Score

Once the propensity modeling is completed, the propensity score is calculated for each patient. The procedure is similar to that used to calculate, for a given patient, expected hospital mortality for coronary artery bypass grafting from the Society of Thoracic Surgeons risk equation.[26]

A logistic regression analysis, such as used for the propensity model, generates a *coefficient* for each variable. The coefficient maps the units of measurement of the variable into units of risk.[1] Specifically, a given patient's value for a variable is transformed into risk units by multiplying it by the coefficient. For example, if the coefficient is 1.13 and the variable is "male" with a value of 1 (for "yes"), the result will be 1.13 risk units. If the coefficient is 0.023 for the variable "age" and a patient is 61.3 years old, 0.023 times 61.3 is 1.41 risk units.

One continues through the list of model variables, multiplying the coefficient by the specific value for each variable. When finished, the resulting products are summed. To this sum is added the *intercept* of the model. The final score is the propensity score. Its units are *logit units*, a word coined by Berkson,[27] formerly of the Mayo Clinic.

Using the Propensity Score for Comparisons

Once the propensity model is constructed and a propensity score is calculated for each patient, 3 common types of comparison are employed: matching, stratification, and multivariable adjustment.

TABLE 3. Comparison of patient characteristics according to long-term aspirin use in matched pairs according solely to propensity score

Patient characteristic	ASA (n = 1351)	No ASA (n = 1351)
Men (%)	49	51
Age (y)	60	61
Smoker (%)	50	50
Resting heart rate (beats/min)	77	76
Ejection fraction (%)	51	51

ASA, Long-term aspirin use.

Matching

The propensity score can be used as the sole criterion for matching pairs of patients.[6,28]

Rarely does one find exact matches. Instead, a patient is selected from the *control* group whose propensity score is nearest to that of a patient in the *case* group. If multiple patients are close in propensity scores, optimal selection among these candidates can be used.[23] Remarkably, problems of matching on multiple variables disappear by compressing "everything known about the patient" into a single score!

Table 3 demonstrates that such matching works astonishingly well. The comparison data sets have all the appearances of a randomized study!

However, unlike a randomized study, the method is unlikely to balance unmeasured variables well. My colleagues and I have built propensity models that purposely exclude variables. When this is done, and the variables excluded are not part of a closely correlated cluster (such as body size), matched pairs differ significantly with respect to these excluded variables. In addition, about 2% to 3% of measured variables, despite being represented in the propensity model, are dissimilar ($P < .05$) in the matched groups. Nevertheless, this is remarkably superior to previous methods of matching.

The average effect of the comparison variable of interest is assessed as the difference in outcome between the groups of matched pairs.

Stratification (Subclassification)

Outcome can be compared within broad groupings of patients, called *strata* or *subclasses,* according to propensity score.[8,10,22] After patients are sorted by propensity score, they are divided into equal-sized groups. For example, they may be split into 5 groups, or quintiles (see Table 2), but fewer or more may be used. Comparison of outcome for the comparison variable of interest is made *within each stratum.*

If a consistent difference in outcome is not observed across strata, intensive investigation is required. Usually, something is discovered about the characteristics of the disease, the patients, or the clinical condition that results in a different outcome.

Multivariable Adjustment

The propensity score for each patient can be included in a multivariable analysis of outcome.[5,7,20] Such an analysis includes *both* the comparison variable of interest *and* the propensity score. The propensity score adjusts the apparent influence of the comparison variable of interest for patient selection differences not accounted for by other variables in the analysis.

Occasionally, the propensity score remains statistically significant in such a multivariable model. This occurrence constitutes evidence that adjustment for selection factors by multivariable analysis alone is ineffective. This does not happen often, but when it does, it is something that cannot be ignored.[3] It may mean that not all variables important for bias reduction have been incorporated into the model, such as when one is using a simple set of variables. It may mean that an important modulating or synergistic effect of the comparison variable occurs across propensity scores as noted above. For example, the mechanism of disease may be different within the quintiles. It may mean that important interactions of the variable of interest with other variables have not have been accounted for, leading to a systematic difference identified by the propensity score.

In some settings in which the number of events is small, the propensity score can be used as the sole means of adjusting for the variable representing the groups being compared.[17]

Get Rid of Oranges?

The propensity score may reveal that a large number of patients in one group do not have scores close to patients in the other.[29] If propensity matching is used, some patients may not be matched. If stratification is used, quintiles of patients may have hardly any matches at one or the other or both ends of the propensity spectrum.

The knee-jerk reaction is to infer that these unmatched patients represent, indeed, apples and oranges, unsuited for direct comparison. Resist the urge to neglect these unmatchable patients![19] The most common reason for lack of matches is that a strong surrogate for the comparison group variable has been included inadvertently in the propensity score (see appendix). This variable must be removed and the propensity model revised.

If this is not the case, the analysis may indeed have identified truly unmatchable cases (mixed fruit). In some settings in which my colleagues and I have observed this phenomenon, it represented a different end of the spectrum of disease for which different therapies had been applied systematically. Often the first clue to this "anomaly" is finding that the influence of the comparison variable of interest is inconsistent across quintiles.

Thus, when apples and oranges and other mixed fruit are revealed by a propensity analysis, investigation should be intensified rather than the oranges simply being set aside. After the investigations are over, comparisons among the well-matched patients can proceed while at the same time the reader can be provided with the boundaries within which a valid comparison was possible.

Limitations, Pitfalls, Alternatives

Randomized Trials

Balancing score methods are not substitutes for properly designed, ethical, randomized clinical trials. They cannot account for unknown variables affecting outcome that are not correlated strongly with measured variables. They lack the discipline and rigor of a randomized trial. Thus, although they constitute the most rigorous methods available for apples-to-apples investigation of causal effects on outcome in the nonrandomized setting, they are not as definitive as randomized trials.

On the other hand, they are more versatile and more widely applicable than randomized trials. For example, one can never randomize whether or not a person will have chronic atrial fibrillation or be a smoker at coronary artery bypass grafting.

Methodologic Issues

Some investigators claim that balancing score methods are valid only for large studies, citing Rubin.[21] It is true that large numbers facilitate certain uses of these scores, such as stratification. Case-control matching is also better when a large group of controls is available for matching. However, I believe that there is considerable latitude in matching that still reduces bias; the method seems to "work," even for modest-sized data sets.

Another limitation is having few variables available for propensity modeling. The propensity score is seriously degraded when important variables influencing selection have not been collected.[2]

The propensity score may not eliminate all selection bias.[30] This may be attributed to limitations of the modeling itself imposed by the linear combination of factors in the regression analysis that generates the balancing score.

Perhaps the most important limitation is inextricable confounding. Suppose one wishes to compare on-pump coronary bypass grafting with off-pump operations. One designs a study to compare the results of institution A, which performs only off-pump bypass, with those of institution B, which performs only on-pump bypass. Even after careful application of propensity score methods, it remains impossible to distinguish between an institutional and a

treatment difference because they are inextricably intertwined—they are the same variable!

Extensions

At times, one may wish to compare more than 2 groups, such as groups representing 3 different valve types. Under this circumstance, multiple propensity models are formulated and used.[21] I prefer to generate fully conditional multiple logistic propensity scores, although some believe this "correctness" is not essential.[31]

Most applications of balancing scores have been concerned with dichotomous (yes/no) comparison group variables. However, balancing scores can be extended to a multiple-state ordered variable (ordinal) or even a continuous variable.[32] An example of the latter is the use of correlates of the continuous value of ejection fraction as a balancing score to isolate the possible causative influence of left ventricular dysfunction.

Conclusions

Be suspicious of apples-to-oranges comparisons! In the past, methods were limited for identifying apples from among the mixed fruit so that a proper comparison could be made. The propensity score and balancing scores in general provide the collaborating statistician with powerful weapons for making valid apples-to-apples comparisons in the nonrandomized or unrandomizable setting. Their theoretical properties and reason for working in this fashion are becoming increasingly clarified, as are their limitations.

I suggest that in settings in which comparison of outcome is based on nonrandomized clinical experience and, therefore, the danger of apples-to-oranges comparison is present, balancing scores should be considered and, if appropriate, used. Because this is my recommendation, you, the reader, need to be "clued in" to this methodology. I hope this explanation has made you a more informed, and less intimidated, reader!

References

1. Kirklin JW, Barratt-Boyes BG. The generation of new knowledge from information, data, and analyses. In: Kirklin JW, Barratt-Boyes BG, editors. Cardiac surgery, Chap 6. New York: Churchill Livingstone; 1993. p. 249-82.
2. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics.* 1993;49:1231-6.
3. Drake C, Fisher L. Prognostic models and the propensity score. *Int J Epidemiol.* 1995;24:183-7.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41-55.
5. Mark DB, Nelson CL, Califf RM, Harrell FE Jr, Lee KL, Jones RH, et al. Continuing evolution of therapy for coronary artery disease: initial results from the era of coronary angioplasty. *Circulation.* 1994;89:2015-25.
6. Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FR Jr, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA.* 1996;276:889-97.
7. Barker FG II, Chang SM, Gutin PH, Malec MK, McDermott MW, Prados MD, et al. Survival and functional status after resection of recurrent glioblastoma multiforme. *Neurosurgery.* 1998;42:709-23.
8. Nakamura Y, Moss AJ, Brown MW, Kinoshita M, Kawai C. Long-term nitrate use may be deleterious in ischemic heart disease: a study using the databases from two large-scale postinfarction studies. *Am Heart J.* 1999;138:577-85.
9. Auerbach AD, Hamel MB, Davis RB, Connors AF, Regueiro C, Desbiens N, et al. Resource use and survival of patients hospitalized with congestive heart failure: differences in care by specialty of the attending physician. *Ann Intern Med.* 2000;132:191-200.
10. Legorreta AP, Leung KM, Berkbigler D, Evans R, Liu X. Outcomes of a population-based asthma management program: quality of life, absenteeism, and utilization. *Ann Allergy Asthma Immunol.* 2000;85:28-34.
11. Feinstein AR. Clinical biostatistics. VII. The rancid sample, the tilted target, and the medical poll-bearer. *Clin Pharmacol Ther.* 1971;12: 134-50.
12. Piantadosi S. Clinical trials: a methodologic perspective. New York: John Wiley; 1997.
13. Loop FD. A surgeon's view of randomized prospective studies. *J Thorac Cardiovasc Surg.* 1979;78:161-5.
14. Schlesselman JJ. Case-control studies. New York: Oxford University Press; 1982.
15. Cologne JB, Shibata Y. Optimal case-control matching in practice. *Epidemiology.* 1995;6:271-5.
16. Rubin DB. Bias reduction using Mahalanobis metric matching. *Biometrics.* 1980;36:393-8.
17. Rosenbaum PR. Optimal matching for observational studies. *J Am Stat Assoc.* 1989;84:1024-32.
18. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol.* 1999;150:327-33.
19. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics.* 1985;41:103-6.
20. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *J Clin Epidemiol.* 1989;42:317-24.
21. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.* 1997;127:757-63.
22. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79:516-24.
23. 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;17:2265-81.
24. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytic approaches. *Ann Rev Public Health.* 2000;21:121-45.
25. Rosenbaum PR. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. *J Am Stat Assoc.* 1984;79:41-8.
26. Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: the Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg.* 1999;67:1205-8.
27. Berkson J. Why I prefer logits to probits. *Biometrics.* 1951;7:327-39.
28. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference. Cary (NC): SAS Institute Inc; 2001, http://www2.sas.com/proceedings/sugi26/p214-26.pdf.
29. Lytle BW, Blackstone EH, Loop FD, Houghtaling PL, Arnold JH, Akhrass R, et al. Two internal thoracic artery grafts are better than one. *J Thorac Cardiovasc Surg.* 1999;117:855-72.
30. Heckman JJ, Ichimura H, Smith J, Todd P. Sources of selection bias in evaluating social programs: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proc Natl Acad Sci U S A.* 1996;93:13416-20.
31. Hosmer DW, Lemeshow S. Polytomous logistic regression. Chap 8. In: Applied logistic regression. New York: John Wiley; 1989. p. 216-38.
32. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics.* 1992;48:479-95.
33. Breiman L. Bagging predictors. *Machine Learning*. 1996;26:123-40.

34. D'Agostino RB Jr, Rubin DB. Estimating and using propensity scores with partially missing data. *J Am Stat Assoc.* 2000;95:749-59.
35. Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ, and the Pneumonia Patient Outcomes Research Team (PORT) Investigators. Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Med Care.* 1995;33:AS56-66.
36. Cook EF, Goldman L. Asymmetric stratification: an outline for an efficient method for controlling confounding in cohort studies. *Am J Epidemiol.* 1988;127:626-39.

Appendix

This appendix is intended for the biostatistical collaborator. It is a "how-we-do-it" (my colleagues and I) commentary, not a mathematical appendix.

Propensity Model Construction

For 2-group comparisons, we construct propensity models with the use of logistic regression. Nearly always it is useful to the investigators and the readers to have a well-formulated explanatory model of the differences between patients receiving one treatment rather than the other. Thus, we begin with parsimonious model construction.

Preparatory analyses. The modeling process involves all the well-known preparatory steps that help one get to know the data in detail. We examine simple correlations (because medical data are inherently redundant), construct contingency tables with respect to the comparison variable of interest, and perform *t* tests for continuous variables. All this is useful not only in screening variables as possible risk factors, but also in eliminating some variables that occur infrequently or are associated with too few events for computational stability. We calibrate continuous and ordinal variables to the event scale by transformation of scale. Only then is multivariable analysis begun.

Explanatory model construction. Variables of good quality, well understood, and appropriate for the analysis are examined without regard to the univariable testing. This means that on occasion, a univariably nonsignificant variable will become significant in the analysis. You will have to investigate whether this is simply an adjusting factor (that may require more work on the main variable), or a variable representing a tiny subset of patients once many variables are in the model, or a lurking variable.

We use a variety of model-building methods. Prominent among these is so-called "bagging" using computer-intensive bootstrapping.[33]

Propensity model construction. However, the propensity model is not parsimonious, but is augmented with whatever is recorded about the patients, and particularly variables that might be related to selection.[22] The object is to account for everything known that may relate to either systematic bias, or simply bad luck, that has otherwise unbalanced the comparison groups of interest.[17] We like to achieve a goodness-of-fit c-statistic in the 0.8 to 0.9 range. Its developers even suggest ignoring usual concerns about model overdetermination. The most useful propensity models incorporate well-calibrated continuous variables so as to produce a smooth distribution of scores.

The one thing *never* considered in forming the propensity model is the outcome of interest. All work must be done without respect to outcome.[4]

A special word is needed about managing missing values for some variables. Because of the high degree of correlation among medical variables, some variables with missing or unreliable values might be ignored. More commonly, methods of imputing missing values, informative or noninformative, should be used. The object is to be able to calculate a propensity score for each patient. We form a set of indicator variables that identify patients who have a missing value for a variable (at least when missing values occur in a substantial number of patients, such as 5% to 10%). These indicator variables are included in the propensity model to distribute missing values appropriately and reduce the bias of missing values.[22,34]

Propensity modeling trap. Beware of variables that are strong surrogates for the group of interest. Some statisticians have remarked that they see no sense in using balancing scores because they already know which patients belong to each group! This reflects lack of understanding of what one is trying to accomplish with the propensity score. (They forget that the same statement can be made about a logistic analysis of hospital mortality.) The object is to produce a model for use in reducing bias of how the patients were selected for the group they are actually in and to permit apples-to-apples comparisons. The danger can be subtle. For example, if the two treatments being compared have been used sequentially in time, then date of treatment (usually a good variable for propensity modeling) is a surrogate for group membership and should not be used.

Despite attention to this detail, quasi-separation in the modeling may occur. One possible explanation for this occurrence is that the variables contain all the information that has actually been used to formulate a rules-based treatment policy. If this is the case, no balancing score will be helpful in evaluating the rules with respect to outcome short of a proper trial.

Alternative models. Just as there are alternatives to logistic regression for analysis of binary outcomes, there are alternatives to its use in forming propensity scores. Thus, any method for classification, such as computer-aided regression trees (CART), neural networks, or optimum discrimination, could be used.[35,36] For some of these methods, it is necessary to dichotomize the explanatory variables, leading to a "lumpy" balancing score that is not ideal.

Calculating the propensity score. One can use the propensity score directly in logit units or convert it to probability. For most uses, it makes no difference. However, it makes a difference if the propensity score is used in a multivariable analysis. In that setting, treat the propensity score as you would any other continuous variable. It may have to be calibrated to the scale of risk by transformation.

Using the Propensity Score for Comparisons by Multivariable Adjustment

As mentioned in the text, the propensity score for each patient can be included in a multivariable analysis of risk factors. We first check that we have a well-matched set of patients, as discussed in the section "Get Rid of Oranges?" in the text. Once a well-matched patient group is available, we have found it useful to first perform an analysis without forcing in the variable of interest or the propensity score. We then look at the variable of interest just as we would do in a randomized trial, this time forcing it into the model and noting which, if any, variables it displaces. We then investigate all the interactions between this variable of interest and the other variables in the model. Finally, we look with equal inten-

sity at the propensity score in the model. This sequence of steps relies heavily on bootstrap bagging.[33]

The sequential strategy described has afforded us the opportunity to better understand the influence on outcome of the comparison variable of interest, as well as the thoroughness of adjustment by risk factors alone. We[29] generally report the magnitude of effect of the comparison variable of interest as the bootstrapped median.

An important consideration is interpretation of a multivariable model when the propensity score remains statistically significant for the multitude of reasons cited in the text. This situation, particularly in a multivariable equation that is intended for prospective prediction, presents an interesting dilemma. All other variables in the model relate to characteristics of individual patients, so they can be applied to a future patient. However, the propensity score represents an attribute of the specific group of patients used in the analysis. A future patient does not belong to this group! Such a mixture of individual and group variables in the same model is an interesting statistical anomaly that is incompletely understood.[3]