Chapter 15.2: Adjusted Analyses in Studies Addressing Therapy and Harm

Thomas Agoritsas; Arnaud Merglen; Nilay D. Shah; Martin O'Donnell; Gordon H. Guyatt

Clinical Scenario

You are an emergency physician working at a secondary care hospital. The institution's board wants to provide the best possible treatments for patients with acute myocardial infarction (MI) and is considering expanding its cardiac catheterization facility to better serve candidates for catheterization.[1] You are requested to assess the evidence regarding that decision.

You are aware of a systematic review of randomized clinical trials (RCTs) comparing a routine invasive treatment (cardiac catheterization) vs conservative therapies (pharmacological therapy, with selective catheterization only for patients with recurrent symptoms or objective inducible ischemia) and showing a relative decrease of 18% in mortality or MI in the more invasively treated group.[2] A colleague argues that in the real world the effect may be much larger, with as much as a 50% relative decrease, as suggested by a cohort study of more than 120 000 Medicare patients hospitalized with MI, some of whom received invasive treatment (ie, catheterization within 30 days) and others medical treatment.[3]

Unsure from where best estimates of the effect of invasive vs conservative therapy should come, you review the report from the large cohort study. You find that the investigators used several adjustment methods: multivariable risk adjustment, propensity score adjustment, propensity-based matching, and instrumental variable analysis (IVA). These analyses showed variable results, ranging from a 16% relative reduction—very similar to that achieved in the RCTs—to a 50% relative decrease in mortality in the more invasively treated group. You want to determine what inferences you can draw from the RCTs as compared with the observational trials that used differing adjustment methods.

Appraising Studies on Therapy and Harm

Large randomized trials ensure that treatment and control groups are balanced with respect to factors associated with outcomes (typically age, sex, disease severity, and comorbidity—"prognostic factors") and thus provide the optimal approach to address questions about the benefits and harms of interventions.[4,5] However, RCTs may not be available or feasible, particularly regarding questions of harm. For example, one may be interested in long-term outcomes years after exposure to an intervention, or in rare serious adverse effects, detection of which will require data from tens of thousands of patients. Randomization also may be unethical, as in the study of risky behaviors. In such instances, the best evidence will come from observational studies. Sophisticated analyses of longitudinal databases are also increasingly cited in the context of comparative effectiveness research. Enthusiasts claim that such studies address limitations in the generalizability of RCTs and thus may provide best estimates of real-world effectiveness.[6-8]

However, observational designs can often produce misleading results, because real-life treatment decisions are typically influenced by patient characteristics likely to be associated with the outcomes of interest (ie, they are likely to be prognostic factors).[9,10] Sophisticated statistical adjustment mechanisms may overcome the problems of prognostic imbalance. The purpose of this Users' Guide is to introduce the fundamental concepts and relative merits of adjustment methods used to account for prognostic imbalance (Table 15.2-1).

Table 15.2-1.

Summary of the Main Approaches to Address Prognostic Imbalance Between Compared Groups

	Description of the Approach	Main Advantages	Main Limitations
Traditional risk-adjusted regression analysis	An analysis that examines the association between treatment selection and the outcome of interest Levels the playing field across prognostic factors between groups by creating prognostically homogenous strata and combining results across strata Multivariable regression simultaneously includes a number of prognostic factors and yields a single estimate of treatment effect adjusted for these factors	Long established and widely used to adjust for confounders in observational studies Interpretation of treatment effects and harms is more straightforward than alternative approaches to adjustment	Investigators may not have measured—or not measured accurately—all relevant prognostic factors Unknown factors may bias the results (ie, residual confounding)
Propensity analysis (propensity-based matching, propensity score adjustment)	A process that computes and assigns to each patient a propensity score, ie, the likelihood of receiving an intervention given that patient's status on measured prognostic factors Investigators then either conduct a regression analysis adjusting for propensity score or choose matched pairs with the very similar propensity score	Can better manage multiple prognostic factors when outcome events are few Can transparently demonstrate prognostic balance within deciles of propensity to receive treatment	As with traditional regression adjustment, does not address residual confounding Empirically, results are often very similar to traditional regression Interpretation is less intuitive
Instrumental variable analysis (IVA)	A method that aims to identify a variable that would substitute for randomization by its association with whether patients did or did not receive the intervention This “instrumental variable” should not be associated with the outcome of interest other than through the intervention under investigation	Theoretically, appropriate instrumental variables have the potential to adjust for unknown prognostic factors, which is not true of either regression or propensity	Appropriate instrumental variables are hard to find; rely on strong assumptions, some of which cannot be empirically verified; are difficult to understand
Randomized clinical trials (RCTs)	An experimental design in which patients are randomly allocated either to the intervention or to the comparator To further minimize risk of bias, RCTs should apply well-established bias-reducing measures (ie, central randomization, allocation concealment, blinding, intention-to-treat analysis, and complete follow-up)	Optimal way to ensure that the intervention and control groups are prognostically balanced Works both for known and unknown prognostic factors	Not always feasible (eg, long-term outcomes, rare adverse events) Not always ethical (eg, risky behaviors)

The Challenge of Prognostic Imbalance

Why Is It Important to Randomize?

Typical answers to this question include to reduce the risk of bias or to create groups that are “similar” or “balanced,” except for the intervention being compared. To grasp what is at stake, consider the type of patient characteristics that should be balanced.[11] Possible characteristics include age, severity of disease (eg, cancer stage), or comorbidities. But how are these factors different from other common characteristics, such as eye or hair color? The key difference is that the former are likely to be associated with patient's prognosis—the likelihood that patients will experience the outcome of interest (eg, mortality or stroke)—and the latter will have no such association.

Ideally, studies evaluating interventions will compare populations identical in all prognostic factors related to the outcome, except for the intervention under study. Whenever feasible, randomization is the optimal way to achieve such balance, not only for known but also for unknown determinants of the outcome.[9,12,13] In contrast, observational studies will invariably involve some prognostic imbalance, because patients who are selected or who self-select for an intervention or an alternative will almost certainly differ with respect to their risk of developing the outcome.[10] A crude analysis of the study will thus produce a biased result, either overestimating or underestimating the true effect. This situation is referred to as *confounding* (and the corresponding factor as a *confounder*[11]) and require an adjusted analysis.[14] Confounders are factors that can affect patients' outcomes, such as age, stage of disease (early vs late), or factors rarely measured, such as the extent of tumor invasion or the degree of adhesions or patient adherence with the treatment.

Even if investigators accurately measure and adjust their analysis for all known prognostic factors, results may still be biased by different distributions of unknown prognostic factors (ie, residual confounding).[12,15,16] Antioxidant supplements for healthy persons represent an example: observational studies demonstrated that individuals consuming supplements had lower rates of all-cause mortality,[17,18] but RCTs revealed that this difference was not related to their regular consumption of supplements. Rather, participants in observational cohorts who took supplements had a more favorable set of prognostic factors (ie, they were healthier to start with). When patients were randomized to receive or not receive the supplements, thus ensuring prognostic balance, the apparent benefit on mortality disappeared.[19,20] This is not an isolated case: a recent meta-epidemiologic survey found that observational studies from routinely collected data showed significantly (31%) more favorable mortality estimates than subsequent RCTs (relative odds ratio [OR], 1.31 [95% CI, 1.03-1.65]).[10] If no RCT is available, clinicians should consider that treatment effects are likely to be smaller—if they exist at all—than those from observational studies.

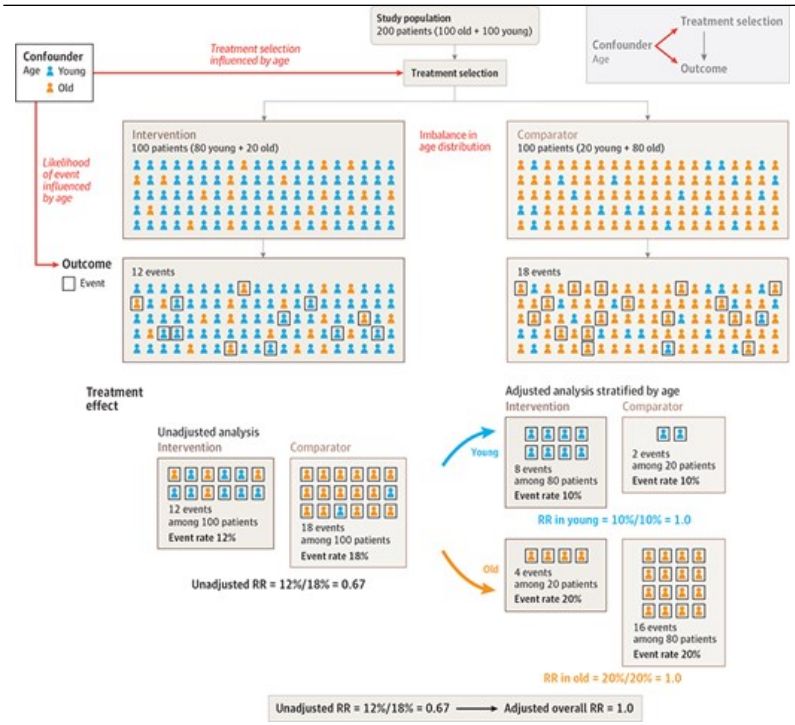
Understanding Stratification and Adjustment

To demonstrate the fundamental logic behind adjustment, consider a hypothetical cohort study including patients with dissimilar ages in the intervention and control group (Figure 15.2-1).[11] Among the 100 patients exposed to the treatment, 80% turned out to be young, while 20% were old. Conversely, among the 100 patients in the comparator group, 80% were old, and 20% were young. The study found 12 deaths in the intervention group (8 young and 4 old patients) and 18 in the control group (2 young and 16 old patients). Overall, the relative risk of dying with the treatment compared with the control was therefore 0.67 (12%/18%). One possible interpretation is that the treatment is effective, resulting in a 33% relative reduction in the outcome of interest. However, age is associated with the risk of death, is not distributed equally in the intervention and control groups, and thus is a likely confounder.

Figure 15.2-1.

Demonstration of Effect of Adjustment for Age on Observed Relative Risk

Example adapted from Kennedy et al.[11] RR indicates relative risk.



One simple way to determine the true benefit is to calculate the effect estimates separately for the 2 prognostic groups, then combine the results across strata (Figure 15.2-1). When restricting the analysis to the young, the mortality rate was 10% in both the intervention group (8/80) and control group (2/20), resulting in a relative risk of 1.0 (10%/10%). As expected, the mortality rates were higher among the old patients but again similar across the intervention (4/20) and control group (16/80), resulting again in a relative risk of 1.0 (20%/20%) (Figure 15.2-1).

Because the intervention group was on average younger and thus at lower risk of dying, the unadjusted analysis spuriously favored the intervention. However, the unconfounded analysis resulted in groups that were more similar to one another than before the stratification, creating 2 comparisons in which groups were homogeneous for the prognostic factor, age. Combining the intervention effects across groups provided an adjusted relative risk of 1.0, showing that the intervention actually had no effect on mortality.[11] Adjustment can also work the other way around, by increasing or revealing an association when the unadjusted analysis showed little or no apparent effect.

Extending the example, consider that the 2 groups also differ in the proportion of patients with diabetes. If that is the case, 4 strata are required to achieve prognostic balance: young and diabetic; young and nondiabetic; old and diabetic; old and nondiabetic. Add another prognostic variable that differs between the groups—presence or absence of hypertension, for instance—and the number of required strata increases to 8. The number of strata will increase further with any additional factors requiring adjustment.

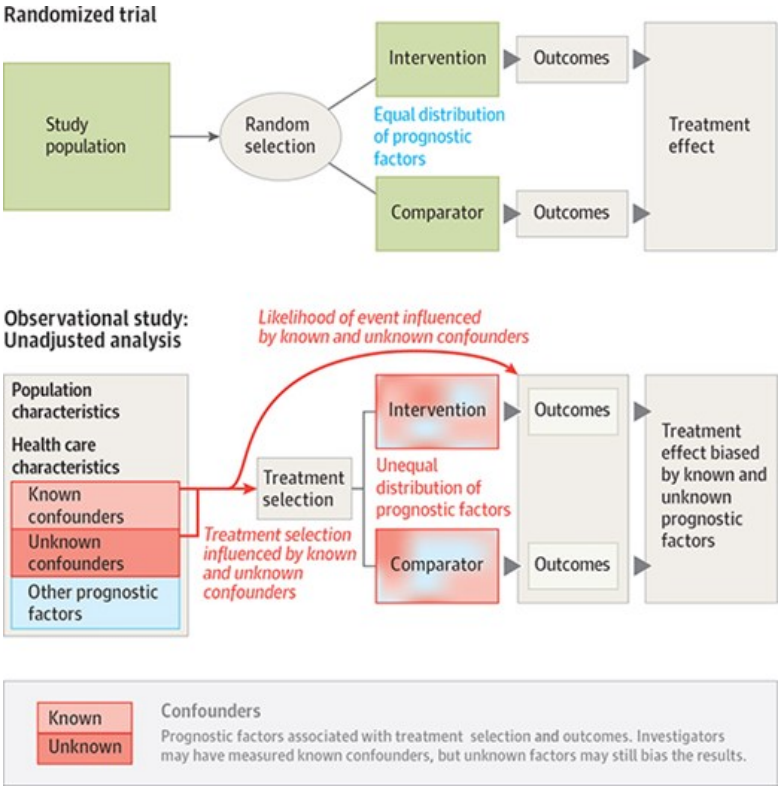
Commonly Used Risk-Adjusted Analysis

When there are multiple prognostic factors to account for, simple stratification becomes impractical to report and interpret, because of the increasing number of strata, and adjustment requires more sophisticated methods, such as multivariable regression analysis. A risk-adjusted regression model is a single equation that predicts the outcome of interest, usually referred to as the dependent variable.[21] The variables that predict or determine the outcome (ie, independent variables) will include the intervention (eg, a drug or surgical intervention) and potential confounders (eg, age, sex, disease severity) that can also influence the outcomes (Figure 15.2-2). Although such analysis requires computer-based computation using statistical software, the adjustment principle remains the same: creating groups that are prognostically balanced, generating estimates in those groups, and combining results across groups. The regression model thus levels the playing field by holding all confounders constant and yields a single estimate of effect that is independent of those confounders.

Figure 15.2-2.

Prognostic Factors in a Study of Therapy or Harm

Prognostic factors become confounders when they are associated with the likelihood that a patient receives an intervention or a comparator and also associated with the likelihood of experiencing the outcome of interest. Known confounding variables are amenable to an adjusted analysis. Unknown confounders are also associated with both the likelihood of receiving the intervention and with experiencing the outcome but are not amenable to an adjusted analysis.



The choice of regression model depends on the nature of the outcome or dependent variable. Dichotomous variables are those that are present or absent (such as mortality or stroke), require logistic regression, and express the effect estimate as the adjusted ORs of the outcome. For instance, an OR of 0.5 for mortality means that the odds of dying in the intervention group are half those in the control group. For continuous variables, such as body weight, systolic blood pressure, or quality-of-life scores, investigators perform adjustment using linear regression, and the effect estimate is an adjusted absolute reduction (eg, so many pounds or kilograms less) between the intervention and control groups.[22] For time-to-event survival analysis —eg, whether an intervention delays the occurrence of an event—adjustment is commonly performed using Cox proportional hazards models, and investigators report results as hazard ratios (HRs).[23]

Users' Guide to a Risk-Adjusted Analysis

Box 15.2-1 summarizes key considerations for users appraising studies addressing therapy or harm that involve statistical adjustment. Questions focus on the initial prognostic imbalance in observational studies, first by asking whether the investigators identified all known prognostic factors for the outcome of interest and second by asking whether these outcomes were accurately measured.

Box 15.2-1.

Users' Guide to an Adjusted Analysis

Did the investigators identify all known prognostic factors for the outcome of interest?
Did the investigators accurately measure all these prognostic factors?
Did the investigators conduct an adjusted analysis that included all these prognostic factors?

Accurate measurement of prognostic variables is important. Administrative databases are often used in observational studies since they can provide large sample sizes, facilitating the study of rare events. However, they are often limited in the quality and depth of their data on patients' characteristics.[14] For instance, in a study assessing the accuracy of electronic reporting compared with manual chart review for patients with diabetes, electronic databases substantially underestimated the provision of appropriate asthma medication and overestimated rates of cholesterol control.[24] Prognostic variable accuracy is usually more problematic when evidence comes from existing databases collected for administrative reasons (eg, billing) rather than collected prospectively to address the clinical question at hand.

The last criterion focuses on the analysis: it is of little use to accurately collect information on prognostic variables unless the information is incorporated in an adjusted analysis. One common question asked regarding regression analysis is how many variables can be included. The answer depends on how many outcome events (such as the number of deaths) occur in the study (ie, the number of events per variable). This technical question is beyond the scope of this Users' Guide and has no definitive answer. However, the more variables investigators adjust for, the more outcome events are required for the regression analysis to function properly. If investigators adjust for too many variables, chance differences in just a few patient observations could lead to very different, and in fact misleading, results.[25-27]

Applying the Guide to the Clinical Scenario

In the cohort study of 122 124 patients who had an MI, about 60% received invasive therapy and 40% conservative therapies. The distributions of key prognostic factors differed substantially between the 2 groups (Table 15.2-2, columns 2 and 3): patients who received cardiac catheterization were younger, more likely male, had lower MI severity (ie, fewer participants with shock and fewer with hypotension), and were more likely admitted to high-volume hospitals. To adjust for this prognostic imbalance, the investigators examined 65 variables potentially associated with post-MI mortality,[3,28] including patient characteristics (eg, age, sex), presentation characteristics (eg, MI location, heart failure), comorbidities (eg, diabetes, hypertension), and hospital characteristics (eg, volume, teaching status). Although this list included a large number of prognostic factors, some may still be missing or not measured, such as alcohol use or chronic inflammatory diseases (eg, chronic obstructive pulmonary disease, rheumatoid arthritis), and some may be suboptimally measured, leading to residual confounding. For instance, although people with the same zip code have very different incomes, the investigators adjusted for socioeconomic level by zip code median social security income. Trained abstractors acquired data from hospital records and used data from Medicare mortality data with follow-up over 7 years to ensure that patient information was as accurate as possible (Box 15.2-1, second criterion).

Table 15.2-2.

Example of the Impact of Propensity Matching on Distributions of Patient Characteristics in a Cohort Study[a]

Characteristic	No. (%)				
	Received Cardiac Catheterization Within 30 Days				Unmatched Patients Receiving Cardiac Catheterization (n = 42 045)
	Overall Cohort		Propensity-Based Matched Cohort		
	No (n = 48	Yes (n =	No (n = 31	Yes (n = 31	

	886)	73 238)	193)	193)	
Predicted 1-y mortality, mean (SD)[b]	32.3 (18.3)	20.9 (13.3)	26.8 (15.5)	27.8 (12.5)	15.8 (7.5)
Demographics					
Age, y					
65-74	40.2	64.4	45.2	45.3	78.6
75-84	59.8	35.6	54.8	54.7	21.4
Men	49.7	58.4	53.2	49.6	65.0
Black race	7.5	4.8	5.7	6.6	3.5
Social Security income≥$2600	30.0	29.7	30.2	30.2	29.2
Comorbidities					
History of angina	44.1	49.9	46.0	45.6	53.2
Previous myocardial infarction	32.9	26.4	28.7	31.9	22.3
Previous revascularization	17.8	20.9	18.0	20.2	21.3
Congestive heart failure	27.2	10.4	16.6	18.3	4.6
Diabetes mellitus	36.6	28.6	31.8	34.1	24.5
Peripheral vascular disease	12.8	9.1	10.6	11.5	7.3
Chronic obstructive pulmonary disease	24.9	17.6	20.9	23.3	13.3
Smoker[c]	16.1	18.0	16.5	17.0	18.8
MI clinical presentation characteristics					
Non–ST-segment elevation acute MI	41.8	38.9	39.8	40.1	38.0
Shock	1.9	1.5	1.8	2.3	0.9
Hypotension	3.5	2.3	3.1	3.6	1.2
Received cardiopulmonary resuscitation	1.8	1.6	2.3	3.5	0.2
Peak creatinine kinase >1000	29.1	32.4	31.7	31.8	32.9

U/L (16.7 µkat/L)					
Annual MI volume >200 patients	20.1	30.4	22.9	20.5	37.8

Abbreviation: MI, acute myocardial infarction.

[a]Adapted from Stukel et al.[3] All data are presented as percentages. Standardized difference is the mean difference divided by the pooled SD, expressed as a percentage.

[b]Predicted 1-year mortality was computed using the Cox proportional hazards regression model, including all baseline patient characteristics of age, sex, race, socioeconomic status, comorbidities, and clinical presentation.

[c]Defined as current smoker.

The investigators compared mortality rates between treatment groups using a Cox regression model and adjusted for all 65 variables. Within the first 4 years, 50 669 patients died. Before adjustment, routine invasive therapy was associated with a 63% relative decrease in mortality (HR, 0.37 [95% CI, 0.36-0.37]). Because those receiving the invasive intervention tended to be lower-risk patients, it was expected that risk adjustment would attenuate the effect estimate: indeed it did, yielding a 49% reduction in risk associated with the invasive intervention (HR, 0.51 [95% CI, 0.50-0.52]). Despite the extensive risk adjustment, one can still suspect that unmeasured factors, such as differences in the prevalence of alcohol use between groups, may bias the estimate of treatment effect (ie, residual confounding).[3,28] Indeed, the relative risk reduction with invasive therapy in the systematic review of RCTs of invasive vs conservative management was 18% (OR, 0.82 [95% CI, 0.72-0.93]),[2] suggesting that the adjusted estimate of an almost 50% relative reduction remains a large overestimate. In other words, because of residual confounding, the adjusted analysis has failed to adequately account for prognostic imbalance.

Propensity Analysis

A Focus on Treatment Selection Bias

When prognostic factors influence clinicians' decision to prescribe—or patients' decision to receive—a given treatment, imbalances in these prognostic factors will bias estimates of treatment effect. Propensity analysis addresses this issue by assigning to each patient a propensity score defined as the likelihood of being exposed to an intervention given that patient's status on measured prognostic factors.[3,29]

How Is the Propensity Score Derived?

Imagine a cancer drug known to have more adverse effects in elderly patients than in younger patients. With the same histological findings, clinicians may prescribe the drug to 60% of young patients but to only 20% of older patients. By accounting for this difference in age-related propensity to receive treatment (ie, 3 times higher in young patients vs older patients), investigators try to address this treatment selection bias.

Furthermore, patients are more likely to receive the drug if they are male, white, and have a low comorbidity score. Investigators will then construct a logistic regression model for the propensity of receiving treatment, with age, sex, race, and comorbidity as the independent variables and receipt of treatment (yes or no) as the dependent variable.

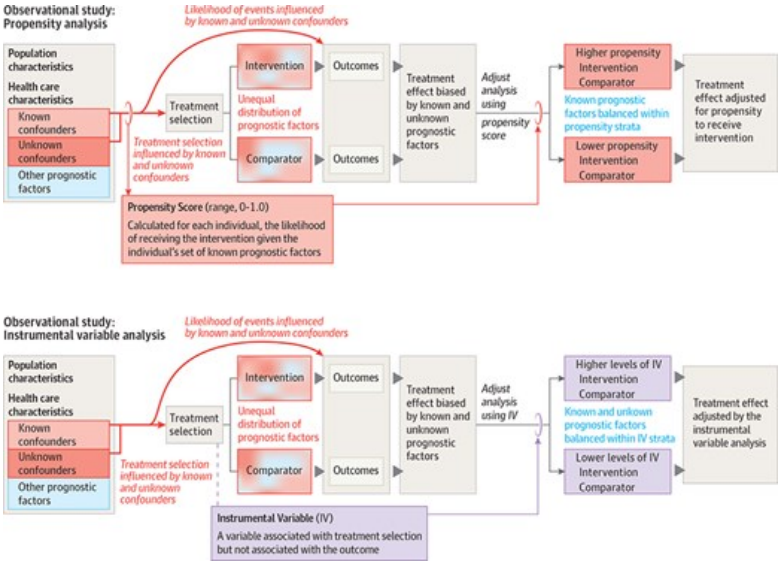
Investigators will then assign to each patient a propensity score according to the patient's likelihood of receiving treatment. The score will range from 0 (no chance of receiving treatment) to 1 (certain to receive treatment).[30-32] An old, female, nonwhite individual with high morbidity will have the lowest propensity score; a young, male, white individual with low comorbidity will have the highest propensity score (Figure 15.2-3).

Figure 15.2-3.

Associations in a Propensity Analysis or an Instrumental Variable Analysis

Top, Prognostic variables associated with the likelihood of receiving the intervention determine the propensity score. Such variables are also

confounding variables because they are associated with the likelihood of occurrence of the outcome of interest. Adjusted analysis can consider the propensity score (red arrow) and potentially other confounders that are not associated with propensity to receive treatment. Models cannot consider unknown confounders. Bottom, Purple arrow denotes that the instrumental variable is associated with whether patients receive or do not receive the intervention of interest. The absence of arrows between the instrumental variable and the confounding variables or the outcome indicates no association with any of these variables.



How Is Propensity Score Used to Adjust for Prognostic Imbalance?

Investigators can use the propensity score, computed as a single continuous variable ranging from 0 to 1, in different types of analyses. First, they can conduct a regression analysis with propensity score and treatment as the independent variables. As for a risk-adjusted regression, this approach attempts to create prognostic variable balance by comparing the outcome between the intervention and the control within groups of patients that are homogeneous (Figure 15.2-3), in this case for their propensity to receive the intervention.[31] A technical refinement of this approach that users will also encounter in reported studies involves weighting patients by the inverse of their propensity score. With this approach, investigators try to achieve prognostic balance by giving less weight to a patient with a higher propensity score and more weight to a patient with a lower propensity to receive treatment. This process is similar to the use of sampling weights in surveys so that they are representative of specific populations.[33-35]

Alternatively, investigators conduct propensity-based matching. For each patient who received the intervention, investigators seek a patient in the comparator group with a similar propensity score, thus simultaneously accounting for all the confounders used to create the propensity score.[30,36-40] Intervention group patients without a match are discarded from the analysis. This process involves technical decisions regarding what value is close enough (is a propensity score of 0.80 sufficiently close to 0.82 to call it a match?) as well as the matching ratio (how many comparison patients to match each patient receiving the intervention?—1:1, 1:2, etc).[37,41] Choosing strict criteria will reduce the number of matched pairs, but less strict criteria may result in residual prognostic imbalances between treatment groups.

Box 15.2-2 summarizes key additional considerations, beyond those for any adjustment, specific to propensity analysis. To build a good propensity score, investigators need to identify all known prognostic factors associated with both the outcome and the likelihood of being exposed to the intervention. Regardless of the adjustment technique used, success of a propensity analysis should be judged on how well it balanced the distribution of pretreatment confounders between the treatment and control groups.[3,32,42,43]

Box 15.2-2.

Users' Guide Specific to a Propensity Analysis

Did the investigators include in the propensity score all known prognostic factors associated with both the outcome and the likelihood of being exposed to the intervention?

Did the investigators measure those prognostic factors accurately?

Was prognostic balance achieved?

Were the distributions of prognostic factor similar across levels of the propensity score or between matched cohorts?

Back to the Clinical Scenario

In the same cohort study of patients with MI, investigators built a score predicting the propensity to receive invasive rather than conservative therapies. They included the same 65 variables associated with outcome, as these may also influence the decision to treat invasively, and developed propensity scores that ranged from 0 to 0.98, with average scores from 0.16 to 0.90 across propensity deciles (each decile being one-tenth of the total number of patients).

As anticipated, 1-year predicted mortality at baseline differed across propensity scores—highest in the lowest propensity decile and progressively lower as propensity increases, suggesting that clinicians are reluctant to pursue invasive strategies in higher-risk patients as estimated by having a propensity closer to 0. However, within each decile, 1-year predicted mortality—reflecting baseline risk using a regression model including baseline age, sex, race, socioeconomic status, comorbidities, and clinical presentation—was similar for invasive and conservative strategies (Table 15.2-3). For example, in the second propensity decile (ie, with a propensity of receiving the invasive treatment ranging from 0.26 to 0.40), 1-year predicted mortality prior to considering the treatment effect was 38.9% for the invasive strategy vs 39.2% for the conservative strategy.

Table 15.2-3.

Example of Balanced Prognosis Across Propensity Score Deciles[a]

	Decile (Range) of Propensity Score									
	1 (0.00-0.26)	2 (0.26-0.40)	3 (0.40-0.50)	4 (0.50-0.58)	5 (0.58-0.65)	6 (0.65-0.70)	7 (0.70-0.75)	8 (0.75-0.80)	9 (0.80-0.85)	10 (0.85-0.98)
No. of Patients										
No cardiac catheterization	10021	8219	6873	5763	4834	3997	3283	2628	2060	1208
Cardiac catheterization	2191	3993	5340	6449	7378	8215	8930	9585	10151	11006
Predicted 1-y Mortality, %[b]										
No cardiac catheterization	54.5	39.2	31.8	27.5	23.4	20.0	17.3	15.3	14.0	13.6
Cardiac catheterization	51.2	38.9	31.8	27.4	23.5	20.0	17.3	15.3	13.5	12.8

[a]Adapted from Stukel et al.[3]

[b]Predicted 1-year mortality was computed using the Cox proportional hazards regression model, including all baseline patient characteristics of age, sex, race, socioeconomic status, comorbidities, and clinical presentation.

Similarly, predicted mortality in the eighth propensity decile (ie, with a propensity of receiving the intervention ranging from 0.75 to 0.80) was exactly the same, 15.3%, in both the invasive strategy and conservative strategy groups. This similar distribution of baseline risk across each decile suggests that propensity matching had created homogeneous prognostic strata, controlling for known confounders. Nevertheless, the results of a regression model adjusting for propensity showed a 46% reduction in relative hazard (HR, 0.54 [95% CI, 0.53-0.55]), still far greater than the reduction from RCTs. This is not surprising, because propensity matching cannot, any more than traditional regression, solve the problem of unmeasured confounders (Figure 15.2-3 and Table 15.2-1).

In a second propensity analysis, investigators matched patients receiving invasive therapy to the closest control whose propensity score differed by less than 10% among those patients within 5 years of age; the matched groups proved very similar for major prognostic features, suggesting success for the propensity match (Table 15.2-2). Of the patients in the invasive therapy group, however, 42 045 had data discarded because investigators could not identify control patients for these individuals, who were younger, had lower MI severity, were more likely admitted to high-volume hospitals, and who almost invariably received catheterization (Table 15.2-2). Because the data were discarded, the comparative effect estimates, if accurate, may not be generalizable to those individuals. Estimates from the propensity-based matching yielded a 47% reduction in hazard (HR, 0.53 [95% CI, 0.51-0.54]), which is far greater than the benefit of the invasive strategy shown in RCTs.

Is Propensity Analysis Superior to Traditional Risk Adjustment?

The primary advantage of propensity analysis over traditional adjustment is that it summarizes treatment selection in a single variable, thus allowing a more straightforward presentation of whether prognostic balance was achieved in matched cohort samples or in deciles of propensity score (see previous discussion of the clinical scenario).[32,42] Another technical advantage of propensity analysis is its suitability when there are fewer events per

adjustment variable. Investigators can indeed adjust for a single variable that summarizes several confounders—the propensity score—in situations in which adjusting for the same confounders one by one may be problematic because outcome events are too few. Compared with traditional regression, this approach actually allows less biased estimates in smaller samples with rare events.[44] Moreover, propensity models are more liberal than traditional regression with the number of factors accounted for, because their success is judged on prognostic balance achieved on the propensity score, rather than on each confounder (a large database can include more than 100 variables in what are called high-dimensional propensity scores).[29,45-47]

Consistent with theoretical mathematical models,[3,42] however, empirical evidence comparing propensity score approaches with multivariate risk adjustment show that results are usually very similar. A systematic review of 43 studies, including 78 exposure-outcome associations,[42] found that 70 showed similar results between multivariate risk adjustment and propensity analysis; only 8 statistically significant associations with regression were not observed with propensity analysis. Propensity matching provided more conservative estimates, but the difference was small—on average, 6.4% closer to finding no difference between the treatments being compared.

In summary, propensity analysis can better deal with multiple prognostic factors when outcome events are few. Propensity analysis does not, however, provide any further protection against residual confounding for unmeasured prognostic variables (Figure 15.2-3).[48,49]

Instrumental Variable Analysis

The problem with traditional observational studies is that the real-world choice of intervention or control—by clinicians, patients, or circumstances—is often by the same clinical characteristics used as prognostic variables in these studies. Thus, there is almost always a substantial imbalance of prognostic variables in observational studies. It would be optimal if a variable could be identified with receiving or not receiving a therapy not associated with any prognostic factors. In theory, such a variable would operate just as randomization does and result in an unbiased comparison. This is the logic underlying IVA, an approach initially developed in econometrics[50] that is becoming more prevalent in comparative effectiveness research.[51,52]

To succeed as a substitute for randomization, IVA must first meet the following conditions: the instrumental variable must be highly associated with the likelihood of being exposed to the type of intervention received, yet not independently influence the outcome, either directly (eg, by influencing quality of care) or by being associated with any known or unknown prognostic factors (Figure 15.2-3).[3,53-55] Box 15.2-3 presents examples of instrumental variables that have been used in practice.

Box 15.2-3.

Examples of Instrumental Variables

A systematic review of use of instrumental variable analysis in studies assessing prescription drug identified 5 types of instrumental variables.[51]
The first were measures of regional variation in health care: if regions vary in rates or accessibility to chemotherapy or cardiac catheterization, region could constitute the randomization equivalent because patients in one region will get one form of care, whereas another type of care is delivered in other regions. This mimics randomization, because the type of care delivered is a function of where a patient receives care and could be completely independent of any clinical features the patient has that might influence treatment decisions.
The second type relates to hospital practice patterns, which also may vary with respect to use of interventions such as cardiac catheterization, but also drug use,[56] based on choices such as drug availability on the hospital formulary.
A third category of instrumental variable could be physician practice, in which preference for drug use (choice of nonsteroidal anti-inflammatory drug, for instance) might differ.
The first 3 choices are analogous to cluster randomized trials in which the unit of randomization might include clinician practices, hospitals, or regions. However, the instrumental variable also may be associated with individual patients, such as elements of their history unrelated to the outcome of interest (eg, history of gout in a study related to influenza vaccination). Calendar time, such as time of approval for a drug, provides another possible instrumental variable.
Last, some studies involve a complex combination of such variables.[57,58]

Appraising the Instrumental Variable

Box 15.2-4 presents a Users' Guide focused on the 3 conditions instrumental variables must meet to succeed as randomization equivalents. First, investigators must report the proportion of patients receiving and not receiving the intervention of interest and show that it substantially differs across status on the instrumental variable. For example, was the drug under consideration used much more frequently when it was on the hospital formulary (the instrumental variable) than when it was not? Empirical measures of association between the instrumental variable and the intervention can be reported in various forms, such as risk differences, ORs, or F statistics.[52,59]

Box 15.2-4.

Users' Guide to an Instrumental Variable Analysis

Is the proposed instrumental variable associated with the likelihood of being exposed to the intervention?
- Did the investigators report on the empirical association?
- Is the magnitude of the association sufficiently strong?

Is it very unlikely that the instrumental variable influences the outcome?
Have investigators demonstrated prognostic balance across the levels of the instrumental variable?

Second, how confident can one be that instrumental variable status does not itself influence outcome? In the example of the drug's availability on the formulary, how confident can users be that hospitals that put the drug on the formulary are not delivering care superior (or inferior) to that delivered by hospitals that do not put the drug on the formulary and that quality of care is not influencing outcome?[53] If that were the case, there is a risk of attributing to the drug effects that are in fact a result of other aspects of care.[59]

Third, investigators must demonstrate that prognostic features are similar across status on the instrumental variable—the equivalent of the demonstration that randomization has done its job of balancing prognosis. Even if the study passes this hurdle, there is another unverifiable assumption that will always leave one more secure with RCTs than any IVA: results will be unbiased only if unmeasured and unknown prognostic factors are similar across instrumental variable groups (Figure 15.2-3).

In practice, reporting of IVA is often suboptimal.[51,52,59] In a systematic survey including 90 studies, only 28% reported appropriate tests of the strength of the associations between instruments and the intervention, and 49% reported associations between the instrumental variables and measured confounders.[52]

How Is IVA Used to Control for Prognostic Imbalance?

Beyond the choice of an appropriate instrument, methodologists are still debating the optimal way of performing IVA.[52,53,59-61] Indeed, more than 15 statistical methods are currently used,[52] and their choice could lead to new sources of bias.[59,62,63] However, to illustrate how one common analysis (called "local average treatment effects") works, compare it with that of an RCT in which all patients randomly allocated to the treatment group (100%), and none allocated to the control group (0%), should receive the treatment. In a way, randomization acts as a perfect instrument, which controls treatment allocation for all patients, and the difference in outcomes between these 2 groups provides an estimate of average treatment effect. In contrast, imagine that a given IVA sorts patients into 2 groups with different treatment rates, one with higher likelihood (eg, 80%) of receiving the treatment and the other with lower likelihood (eg, 20%), thus functioning partly as a randomizer.[54] Using various statistical techniques, the IVA estimates the treatment effect on the 60% of patients (80% minus 20%) in whom variations in the instrument actually affected the treatment rates and hence the likelihood of being treated (whereas randomization affects treatment allocation for all patients in a trial).[54,56,64] In doing so, IVA often produces effect estimates on an absolute rather than a relative scale.[3]

Could one expect differences between IVA and more traditional risk-adjusted analysis? The answer is yes, at least sometimes.[54,56] First, estimates of treatment effect may be less precise, and thus less likely to be significant, because a systematic review of IVA showed they often resulted in wider confidence intervals than traditional adjustment.[52,65] Second, IVA might indeed adjust for unmeasured confounding and provide estimates closer to the truth. However, even if an IVA meets all of the criteria in this Users' Guide, the question of whether it has really overcome the problem of unknown prognostic imbalance, and further whether it introduces new prognostic imbalance if outcome is associated with the instrumental variable, remains unanswerable[66,67]—until investigators have conducted the corresponding RCTs providing more definitive answers.

Back to the Clinical Scenario

The investigators suggested that regional variation in 30-day cardiac catheterization rates across 566 coronary angiography service areas could serve as instrumental variable to address prognostic imbalance between routine invasive therapy and conservative therapies.[3] The analysis meets the first criterion: cardiac catheterization rates ranged from 29% to 82% across regions and on average 43% to 65% across regional catheterization quintiles (Table 15.2-4), demonstrating a sufficiently strong association between the instrumental variable and the intervention and suggesting that the likelihood that a patient will receive cardiac catheterization was strongly dependent on what hospital they were in, because that hospital had a higher likelihood of performing cardiac catheterization, irrespective of the patient's clinical condition. This assumption is supported by the observation that prognostic factors, including predicted mortality from these factors, were very similar across catheterization quintiles (Table 15.2-4).

Table 15.2-4.

Example of Instrumental Variable Balancing Prognosis[a]

	Quintile (Range) of Regional Cardiac Catheterization Rate, %				
	1 (29.2-48.1)	2 (48.2-53.0)	3 (53.1-56.3)	4 (56.4-60.2)	5 (60.3-82.3)
No. of patients	24872	24184	24718	24063	24287
Mean predicted 1-y mortality, mean (SD), %[b]	26.1	26.0	25.5	25.3	24.6
Demographics					

Age, y					
65-74	53.3	54.4	54.6	55.6	55.6
75-84	46.7	45.6	45.4	44.4	44.4
Men	53.7	54.2	55.0	55.6	56.4
Black race	4.1	8.1	6.3	5.5	5.4
Social Security income≥$2600	30.4	28.2	33.4	27.9	29.1
Comorbidities					
History of angina	50.1	48.3	47.8	47.6	44.0
Previous MI	30.1	29.8	29.2	28.7	26.9
Previous revascularization	16.5	18.6	20.8	20.2	22.1
Congestive heart failure	18.4	18.0	17.3	16.9	15.1
Diabetes mellitus	32.9	32.5	32.3	31.3	30.0
Peripheral vascular disease	10.5	10.9	11.0	10.4	10.0
Chronic obstructive pulmonary disease	21.1	20.2	20.3	20.3	20.7
Smoker[c]	16.7	16.7	17.0	18.0	17.9
MI clinical presentation characteristics					
Non–ST-segment elevation acute MI	40.4	41.2	40.5	39.3	39.0
Shock	1.6	1.6	1.6	1.7	1.7
Hypotension	2.8	2.9	2.6	2.8	2.7
Received cardiopulmonary resuscitation	1.6	1.7	1.7	1.8	1.7
Peak creatinine kinase >1000 U/L (16.7 µkat/L)	30.3	30.5	30.4	31.7	32.6

Abbreviation: MI, myocardial infarction.

[a]Adapted from Stukel et al.[3]

[b]Predicted 1-year mortality was computed using the Cox proportional hazards regression model, including all baseline patient characteristics of age, sex, race, socioeconomic status, comorbidities, and clinical presentation. Standard deviation for predicted 1-year mortality was 16.3.

[c]Defined as current smoker.

However, region might affect mortality aside from the use of cardiac catheterization. High-use regions may have higher volumes, specialized staff,

superior equipment, and superior quality of care, all of which could improve outcomes that one would then falsely attribute to higher rates of invasive management.[3,43] The association between unmeasured confounders and the outcome also cannot be verified.

The results, however, may leave one ready to dismiss theoretical concerns about the degree with which IVA assumptions are actually met. The adjusted IVA estimated a 9.7% absolute mortality reduction associated with the invasive strategy analysis, which the authors converted to a 16% relative decrease in mortality over the 7-year follow-up (HR, 0.84 [95% CI, 0.79-0.90]). This contrasts with the 50% relative decrease observed with both risk-adjusted and propensity-matched analyses but approximates the 18% reduction in hazard from RCTs and suggests that seeking care in a particular region has operated as the equivalent of random allocation to invasive or conservative management strategies. Without the comparison with RCT data, which can suggest the magnitude of residual confounding, believing the IVA results rather than standard regression or propensity analysis would be considerably less secure (unless one was ready to assume that large treatment effects are seldom true for any study design—perhaps not an unreasonable assumption).

Conclusions

Traditional risk adjustment in observational studies is vulnerable to differences in unmeasured or unknown prognostic factors between groups (residual confounding) that leave the results open to bias. Propensity analysis has advantages over traditional regression-based risk adjustment, but the advantages are small and provide little, if any, added protection against residual confounding.[9] IVA has the theoretical potential of also accounting for unmeasured confounders, and some results suggest that it may sometimes do so when used appropriately. However, IVA relies on strong assumptions, some of which cannot be empirically verified. Comparative effectiveness research relying on observational studies using conventional or novel adjustment procedures risks providing the misleading effect estimates seen with hormone replacement for cardiovascular risk,[68,69] β-blockers for mortality in noncardiac surgery,[70] antioxidant supplements for healthy people,[19,20] and statins for cancer.[71,72] If RCTs cannot be conducted, it will remain impossible to determine whether adjusted estimates are accurate or misleading.[9,10,12]

Acknowledgment

The following disclosures were reported at the time this original article was first published in *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

References

1. Rathore SS, Epstein AJ, Volpp KG, Krumholz HM Regionalization of care for acute coronary syndromes. *JAMA*. 2005;293(11):1383-–1387. [PubMed: 15769973]

2. Mehta SR, Cannon CP, Fox KA, et al. Routine vs selective invasive strategies in patients with acute coronary syndromes. *JAMA*. 2005;293(23):2908-–2917. [PubMed: 15956636]

3. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias. *JAMA*. 2007;297(3):278-–285. [PubMed: 17227979]

4. Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care. *JAMA Intern Med*. 2014;174(5):710-–718. [PubMed: 24663331]

5. Walsh M, Perkovic V, Manns B et al. Therapy (randomized trials). In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature*. 3rd ed. New York, NY: McGraw-Hill; 2015.

6. Sullivan P, Goldmann D The promise of comparative effectiveness research. *JAMA*. 2011;305(4):400-–401. [PubMed: 21266687]

7. Hochman M, McCormick D Characteristics of published comparative effectiveness studies of medications. *JAMA*. 2010;303(10):951-–958. [PubMed: 20215609]

8. Demaria AN Comparative effectiveness research. *J Am Coll Cardiol*. 2009;53(11):973-–975. [PubMed: 19281929]

9. Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? *JAMA*. 2014;312(2):129-–130. [PubMed: 25005647]

10. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials. *BMJ*. 2016;352:i493. [PubMed: 26858277]

11. Kennedy CC, Jaeschke R, Keitz S, et al. Tips for teachers of evidence-based medicine: adjusting for prognostic imbalances (confounding variables) in studies on therapy or harm. *J Gen Intern Med*. 2008;23(3):337-–343. [PubMed: 18175191]

12. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286(7):821-–830. [PubMed: 11497536]

13. Saquib N, Saquib J, Ioannidis JP Practices and impact of primary outcome adjustment in randomized controlled trials. *BMJ*. 2013;347:f4313. [PubMed: 23851720]

14. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V Users' Guides to the Medical Literature, IV: how to use an article about harm. *JAMA*. 1994;271(20):1615-–1619. [PubMed: 8182815]

15. Brignardello-Petersen R, Ioannidis JPA, Tomlinson G, Guyatt G Surprising results of randomized trials. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature*. 3rd ed. New York, NY: McGraw-Hill; 2015.

16. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-–926. [PubMed: 18436948]

17. Stanner SA, Hughes J, Kelly CN, Buttriss J A review of the epidemiological evidence for the "antioxidant hypothesis". *Public Health Nutr*. 2004;7(3):407-–422. [PubMed: 15153272]

18. Willcox JK, Ash SL, Catignani GL Antioxidants and prevention of chronic disease. *Crit Rev Food Sci Nutr*. 2004;44(4):275-–295. [PubMed: 15462130]

19. Bjelakovic G, Nikolova D, Gluud C Antioxidant supplements to prevent mortality. *JAMA*. 2013;310(11):1178-–1179. [PubMed: 24045742]

20. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database Syst Rev*. 2012;3(3):CD007176.

21. Ebrahim S, Walter SD, Cook DJ, Jaeschke R, Guyatt G Correlation and regression. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature*. 3rd ed. New York, NY: McGraw-Hill; 2015.

22. Guyatt GH, Haynes RB, Sackett DL Analyzing data. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, eds. *Clinical Epidemiology: How to Do Clinical Practice Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2006.

23. Kleinbaum DG, Klein M *Survival Analysis: A Self-Learning Text*. 3rd ed. New York, NY: Springer; 2012.

24. Kern LM, Malhotra S, Barrón Y et al. Accuracy of electronically reported "meaningful use" clinical quality measures. *Ann Intern Med*. 2013;158(2):77-–83. [PubMed: 23318309]

25. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV Performance of logistic regression modeling. *J Clin Epidemiol*. 2011;64(9):993-–1000. [PubMed: 21411281]

26. Guyatt GH Determining prognosis and creating clinical decision rules. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P eds. *Clinical Epidemiology: How to Do Clinical Practice Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2006.

27. Austin PC, Steyerberg EW The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol*. 2015;68(6):627--636. [PubMed: 25704724]

28. Stukel TA, Lucas FL, Wennberg DE Long-term outcomes of regional variations in intensity of invasive vs medical management of Medicare patients with acute myocardial infarction. *JAMA*. 2005;293(11):1329--1337. [PubMed: 15769966]

29. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516--524.

30. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399--424. [PubMed: 21818162]

31. Austin PC. A tutorial and case study in propensity score analysis. *Multivariate Behav Res*. 2011;46(1):119--151. [PubMed: 22287812]

32. Biondi-Zoccai G, Romagnoli E, Agostoni P, et al. Are propensity scores really superior to standard multivariable analysis? *Contemp Clin Trials*. 2011;32(5):731--740. [PubMed: 21616172]

33. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262--270. [PubMed: 16371515]

34. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137--2148. [PubMed: 20108233]

35. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes [published online April 30, 2015]. *Stat Methods Med Res*. doi: 10.1177/0962280215584401

36. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006. *J Thorac Cardiovasc Surg*. 2007;134(5):1128--1135. [PubMed: 17976439]

37. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150--161. [PubMed: 20925139]

38. Crown WH. Propensity-score matching in economic analyses. *Appl Health Econ Health Policy*. 2014;12(1):7--18. [PubMed: 24399360]

39. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. *J Clin Epidemiol*. 2013;66(11):1302--1307. [PubMed: 23972521]

40. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal. *J Clin Epidemiol*. 2015;68(2):112--121. [PubMed: 25433444]

41. Luo Z, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research. *Med Care Res Rev*. 2010;67(5):528--554. [PubMed: 20442340]

42. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies. *J Clin Epidemiol*. 2005;58(6):550--559. [PubMed: 15878468]

43. D'Agostino RB, Agostino RB Sr. Estimating treatment effects using observational data. *JAMA*. 2007;297(3):314--316. [PubMed: 17227985]

44. 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. [PubMed: 12882951]

45. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol*. 2011;173(12):1404--1413. [PubMed: 21602301]

46. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512--522. [PubMed: 19487948]

47. Martens EP, Pestman WR, de Boer A, et al. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *Int J Epidemiol*. 2008;37(5):1142--1147. [PubMed: 18453634]

48. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? *Eur Heart J*. 2012;33(15):1893--1901. [PubMed: 22711757]

49. Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. *J Clin Epidemiol*. 2014;67(8):932--939. [PubMed: 24774469]

50. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health*. 1998;19:17--34. [PubMed: 9611610]

51. Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. *J Clin Epidemiol*. 2011;64(6):687--700. [PubMed: 21163621]

52. Davies NM, Smith GD, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies. *Epidemiology*. 2013;24(3):363--369. [PubMed: 23532055]

53. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537--554. [PubMed: 20354968]

54. Fang G, Brooks JM, Chrischilles EA. Apples and oranges? interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data. *Am J Epidemiol*. 2012;175(1):60--65. [PubMed: 22085626]

55. Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. *Am J Med*. 2010;123(12)(suppl 1):e16--e23. [PubMed: 21184862]

56. Fang G, Brooks JM, Chrischilles EA. Comparison of instrumental variable analysis using a new instrument with risk adjustment methods to reduce confounding by indication. *Am J Epidemiol*. 2012;175(11):1142--1151. [PubMed: 22510277]

57. Lee CC, Ho HC, Hsiao SH, et al. Infectious complications in head and neck cancer patients treated with cetuximab: propensity score and instrumental variable analysis. *PLoS One*. 2012;7(11):e50163. [PubMed: 23209663]

58. Lalani T, Cabell CH, Benjamin DK, et al. Analysis of the impact of early surgery on in-hospital mortality of native valve endocarditis: use of propensity score and instrumental variable methods to adjust for treatment-selection bias. *Circulation*. 2010;121(8):1005--1013. [PubMed: 20159831]

59. Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology*. 2013;24(3):370--374. [PubMed: 23549180]

60. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects. *Int J Biostat*. 2007;3(1):14.

61. Swanson SA, Hernán MA. Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. *Stat Sci*.

2014;29(3):371-–374. [PubMed: 25580054]

62. Swanson SA, Robins JM, Miller M, Hernán MA. Selecting on treatment: a pervasive form of bias in instrumental variable analyses. *Am J Epidemiol*. 2015;181(3):191-–197. [PubMed: 25609096]

63. Martens EP, Pestman WR, de Boer A, et al. Instrumental variables: application and limitations. *Epidemiology*. 2006;17(3):260-–267. [PubMed: 16617274]

64. Tan HJ, Norton EC, Ye Z, et al. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. *JAMA*. 2012;307(15):1629-–1635. [PubMed: 22511691]

65. Pirracchio R, Sprung C, Payen D, Chevret S. Benefits of ICU admission in critically ill patients: whether instrumental variable methods or propensity scores should be used. *BMC Med Res Methodol*. 2011;11:132. [PubMed: 21936926]

66. Crown WH, Henk HJ, Vanness DJ. Some cautions on the use of instrumental variables estimators in outcomes research. *Value Health*. 2011;14(8):1078-–1084. [PubMed: 22152177]

67. Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Ann Intern Med*. 2014;161(2):131-–138. [PubMed: 25023252]

68. Finucane FF, Madans JH, Bush TL, et al. Decreased risk of stroke among postmenopausal hormone users. *Arch Intern Med*. 1993;153(1):73-–79. [PubMed: 8422201]

69. Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women: the Women's Health Initiative. *JAMA*. 2003;289(20):2673-–2684. [PubMed: 12771114]

70. Devereaux PJ, Yang H, Yusuf S, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial). *Lancet*. 2008;371(9627):1839-–1847. [PubMed: 18479744]

71. Blais L, Desgagné A, LeLorier J. 3-Hydroxy-3-methylglutaryl coenzyme A reductase inhibitors and the risk of cancer. *Arch Intern Med*. 2000;160(15):2363-–2368. [PubMed: 10927735]

72. Dale KM, Coleman CI, Henyan NN, Kluger J, White CM. Statins and cancer risk: a meta-analysis. *JAMA*. 2006;295(1):74-–80. [PubMed: 16391219]