The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 6

A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies: A Review

Peter C. Austin, *Institute for Clinical Evaluative Sciences*
Andreas Laupacis, *St. Michael's Hospital*

Recommended Citation:
Austin, Peter C. and Laupacis, Andreas (2011) "A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies: A Review," *The International Journal of Biostatistics*: Vol. 7 : Iss. 1, Article 6.
Available at: http://www.bepress.com/ijb/vol7/iss1/6
DOI: 10.2202/1557-4679.1285

©2011 Berkeley Electronic Press. All rights reserved.

A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies: A Review

Peter C. Austin and Andreas Laupacis

Abstract

In randomized controlled trials (RCTs), treatment assignment is unconfounded with baseline covariates, allowing outcomes to be directly compared between treatment arms. When outcomes are binary, the effect of treatment can be summarized using relative risks, absolute risk reductions and the number needed to treat (NNT). When outcomes are time-to-event in nature, the effect of treatment on the absolute reduction of the risk of an event occurring within a specified duration of follow-up and the associated NNT can be estimated. In observational studies of the effect of treatments on health outcomes, treatment is frequently confounded with baseline covariates. Regression adjustment is commonly used to estimate the adjusted effect of treatment on outcomes. We highlight several limitations of measures of treatment effect that are directly obtained from regression models. We illustrate how both regression-based approaches and propensity-score based approaches allow one to estimate the same measures of treatment effect as those that are commonly reported in RCTs. The CONSORT statement recommends that both relative and absolute measures of treatment effects be reported for RCTs with dichotomous outcomes. The methods described in this paper will allow for similar reporting in observational studies.

KEYWORDS: randomized controlled trials, observational studies, causal effects, treatment effects, absolute risk reduction, relative risk reduction, number needed to treat, odds ratio, survival time, propensity score, propensity-score matching, regression, non-randomized studies, confounding

Author Notes: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported in part by a Career Investigator Award from the Heart and Stroke Foundation of Ontario. The data used in this study were obtained from the EFFECT study. The EFFECT study was funded by a Canadian Institutes of Health Research (CIHR) Team Grant in Cardiovascular Outcomes Research.

1. Introduction

Randomized controlled trials (RCTs) are considered the gold standard for estimating the efficacy and safety of health care interventions. In RCTs, randomized treatment assignment ensures that, on average, treated and untreated subjects do not differ systematically from one another. Since treatment is not confounded with baseline covariates, simple measures of treatment effect can be estimated by comparing outcomes directly between treated and untreated subjects.

There is a growing interest in using observational studies to estimate the effects of treatments, interventions, and exposures on health outcomes. In observational studies, treatment assignment is not assigned at random, but is influenced by patient characteristics. Observational studies allow one to examine the effects of treatments in settings in which randomized trials may not be feasible; furthermore, they allow one to estimate treatment safety and efficacy outside the tightly controlled confines of an RCT. Since treatment is not assigned at random, treated and untreated subjects frequently differ systematically from one another. Since treatment is confounded with baseline covariates, the effect of treatment on outcomes cannot be estimated by directly comparing outcomes between treated and untreated subjects. Instead, regression analysis is frequently used to estimate the effect of treatment on outcomes after adjusting for differences in baseline characteristics between treated and untreated subjects. A limitation to this approach is that measures of treatment effect obtained directly from regression models are usually in a different metric from those obtained directly from an RCT. For instance, when outcomes are binary, regression adjustment allows for the estimation of an adjusted odds ratio; in an RCT, one can estimate not only the odds ratio, but also the relative risk reduction, the absolute risk reduction and the numbers needed to treat (NNT).

Patients, clinicians and policy makers are usually more interested in understanding the absolute benefits and harms of a therapy, than its relative benefits and harms. For example, although oral contraceptives increase the relative likelihood of a deep venous thrombosis or pulmonary embolism by up to three times (Douketis, 1997), physicians are comfortable recommending an oral contraceptive to most women wishing to avoid pregnancy because the absolute risk of serious harm is extremely low.

An advantage of the relative risk reduction is that its magnitude is often similar in different subgroups of patients. For example, warfarin (an oral blood thinner) decreases the relative risk of stroke in patients with atrial fibrillation (a common heart rhythm disorder that is responsible for a high proportion of strokes in elderly persons) by about 67 percent, irrespective of patient age (Atrial Fibrillation Investigators, 1994; van Walraven, 2009). However, the risk of stroke in people with atrial fibrillation varies greatly, depending upon age and the

presence or absence of other conditions. A young person with atrial fibrillation who has no other cardiac disease has an annual risk of stroke of approximately 1%, while an elderly person with heart failure and a previous stoke has an annual risk of stroke of approximately 15% (Gage, 2001). Since the relative risk reduction of 67 percent appears to be constant across subgroups (van Walraven, 2009), this means that the absolute decrease in stroke associated with one year of warfarin therapy is approximately 0.67 percent for the young person and 10% for the elderly person. The corresponding numbers needed for one year of warfarin therapy are 149 and 10 respectively. Because warfarin therapy is associated with an annual risk of a serious bleeding event (e.g. bleeding into the brain or the stomach) of about 1-2 percent, most clinicians would not recommend warfarin therapy to the young person, but feel that the benefits of warfarin therapy outweigh the risks in the older person. Current guidelines for the management of patients with atrial fibrillation, which influence hundreds of thousands of patients around the world, recommend warfarin for patients at high risk of stroke, but aspirin for those at low risk (Singer, 2008). This example illustrates that, even in the presence of a constant relative risk reduction, differences in patients' baseline risk can markedly impact the clinical impact of a therapy. When describing the benefits and harms of a therapy, it is important to use both relative and absolute expressions of risk reduction (or increase). Furthermore, to improve clinical decision making, it is important to have an understanding of the baseline risk of an event occurring in the absence of treatment.

The objective of this paper is to describe how clinically-meaningful measures of treatment effect can be estimated in observational studies. In particular, we demonstrate how all the measures of effect commonly reported in RCTs can also derived for observational studies. The paper is structured as follows. In Section 2, we describe commonly reported measures of treatment effect in RCTs. We focus on measures of effect when outcomes are continuous, binary, and time-to-event in nature. In Section 3, we introduce the data that will be used throughout the paper to illustrate the different statistical methods. In Section 4, we describe conventional measures of effect obtained in observational studies when regression adjustment is used to account for systematic differences between treated and untreated subjects. Limitations with these measures of effect are highlighted. In Section 5, we describe regression-based approaches to estimating clinically-meaningful measures of treatment effect similar to those obtained in RCTs. In Section 6, we describe methods based on the propensity score that allow for estimating measures of effect similar to those reported in RCTs. Finally, in Section 7, we summarize our observations.

DOI: 10.2202/1557-4679.1285

2. Measures of treatment effects in RCTs

In this section, we review measures of treatment effect that are commonly reported in RCTs and summarize clinical commentaries on the relative utility of these different measures of effect. While we describe on measures of treatment effect when outcomes are continuous, binary, or time-to-event in nature our primary focus is on binary or time-to-event outcomes, as they occur more frequently in published reports of RCTs in the medical literature (Austin et al., 2010a).

2.1. Continuous outcomes

When outcomes are continuous, the difference in mean outcomes between treatment groups is commonly reported in RCTs (Austin et al., 2010a). The treatment effect is the difference in means: the mean change in response due to treatment. In some instances, when the outcome variable also exists as a baseline variable, investigators can use regression models to estimate the effect of treatment on the outcome response after adjusting for the baseline value of the response variable (Senn,1989; Senn, 1994; Altman and Dore, 1991; Permutt, 1990). Researchers can also estimate the effect of treatment on the change in the response variable from the baseline value of the response variable.

2.2. Binary or dichotomous outcomes

There are several possible measures of treatment effect when outcomes are binary in nature. The effect of treatment on outcomes can be reported using risk differences (or absolute risk reductions), the number needed to treat (NNT), the relative risk (or the relative risk reduction), and the odds ratio. Consider a two-armed RCT with a binary outcome Y. Let p_T = Pr(Y=1|Treated) and p_C = Pr(Y=1|Control) denote the probabilities of the outcome (success/failure) in the treated and untreated arms, respectively. Possible measures of treatment effect are described in the table below.

Measure of effect	Definition
Absolute risk reduction (ARR)	$p_C - p_T$
Number needed to treat	1/ARR
Relative risk	p_T/p_C
Relative risk reduction	$100 \times \frac{p_C - p_T}{p_C}$
Odds ratio	$\frac{p_T/(1-p_T)}{p_C/(1-p_C)}$

As illustrated with the previous example of warfarin therapy in patients with atrial fibrillation, the ARR and NNT provide more clinically useful information to patients and clinicians than does the RRR on its own. However, the RRR is needed in order to calculate the ARR and NNT for patients who have different pre-treatment risks of a bad outcome (stroke, in the atrial fibrillation example). Therefore, many recommend that the benefits and harms of therapies should be expressed in both relative and absolute terms. For example, the recently revised CONSORT statement recommends that, for RCTs with dichotomous outcomes, both relative and absolute measures of treatment effect be reported (Schulz et al., 2010). Furthermore, the *BMJ* (*British Medical Journal*) requires that absolute risk reductions and the associated number needed to treat (NNT) be reported for any randomized controlled trial with a dichotomous outcome (http://resources.bmj.com/bmj/authors/types-of-article/research. Site accessed April 5, 2010).

2.3. Time-to-event or survival outcomes

Kaplan-Meier survival curves are frequently reported in RCTs in which outcomes are time-to-event in nature (Austin et al., 2010a). From the Kaplan-Meier survival curves, one can determine the probability of an event occurring within a specified duration of follow-up. From this quantity, one can determine the absolute reduction in the probability of an event occurring within a specific duration of follow-up. In addition, one can estimate the NNT to avoid one event from occurring within the specified duration of follow-up. The area under the survival curve is the expected (or mean) survival time in a given treatment group (Klein and Moeschberger, 1997). Therefore, one can estimate the expected survival for treated and untreated subjects, separately. The difference in expected survival time is the effect of treatment on expected survival. Finally, a Cox regression model can be fit, allowing one to estimate the effect of treatment on the relative hazard of the event occurring (Cox and Oakes, 1984). The first two measures of

DOI: 10.2202/1557-4679.1285

treatment effect are absolute measures of effect, while the third is a relative measure of effect.

There is a paucity of discussion about the relative utility of different measures of treatment effect in RCTs when the outcome is time-to-event in nature. However, many of the arguments in the context of binary outcomes would hold when outcomes are time-to-event in nature. Therefore, at the very minimum, absolute measures of treatment effect should complement relative measures of treatment effect.

3. Dataset for illustrating statistical methods

In this section we briefly describe the data that were used for illustrating the different statistical methods for quantifying the effect of treatment on outcomes. Detailed clinical data obtained by retrospective chart review were available on a sample of 9,945 patients hospitalized with a diagnosis of heart failure from 103 acute care hospitals in Ontario, Canada between April 1, 1999 and March 31, 2001. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, an ongoing initiative to improve the quality of care for patients with cardiovascular disease in Ontario (Tu et al., 2004, 2009). Data on patient demographics, vital signs at presentation, and results of physical examination at presentation, medical history, and results of laboratory tests were collected for this sample. We restricted the initial sample to those 9,162 patients discharged alive with a diagnosis of heart failure. Subjects with missing data on key continuous baseline covariates were excluded from the current study (N = 1,547). We chose to conduct a complete case analysis and exclude subjects with missing data since the focus of the current article is on describing methods to estimate clinically- and policy-relevant measures of treatment effect. In a particular application, applied researchers would need to decide how best to account for missing data. The selected approach could possibly include imputing missing data, thereby allowing for all subjects to be included in the subsequent analyses. Furthermore, the hierarchical nature of the data was not taken into account when estimating different measures of treatment effect. The focus of the paper is on methods to estimate clinically and policy-relevant measures of treatment effect. Consequentially, we did not focus on methods to incorporate the

effect of clustered data into these analyses. Post-discharge mortality was determined by linking the study sample to the Registered Persons Database using encrypted versions of the each patient's Ontario Health Insurance Plan number. A further two subjects for whom the recorded date of death preceded the date of hospital admission were excluded. Thus, the resultant cohort consisted of 7,613 patients. In the current study, the exposure was receipt of a prescription for a beta-blocker at hospital discharge. From this sample of 7,613 patients, we excluded 117 subjects with either a documented history of allergy to beta-blockers or a documented intolerance to beta-blocker use while in hospital (the most common documented reasons for in-hospital intolerance: allergy/hypersensitivity; asthma/COPD/bronchospasm; bradycardia/heart block; hypotension). Therefore, the final study sample consisted of 7,496 patients.

The demographic and clinical characteristics of beta-blocker users and non-users are described in Table 1 (this table is reproduced from a prior study using these data for illustrative purposes (Austin, 2009a)). Continuous variables were compared between treatment groups using the Kruskal-Wallis test, while categorical variables were compared using the Chi-squared test. 27.2% of patients received a prescription for a beta-blocker at hospital discharge. Systematic differences in several baseline covariates were observed between treated and untreated subjects. As noted above, unless other stated, conventional statistical analyses conducted in this sample assume that all observations were independent of one another.

In the current study, we considered two different outcomes: death within one year of hospital discharge (a dichotomous outcome) and time to death, with subjects censored after five years of follow-up (a survival outcome). In our sample, 27.7% of the subjects died within one year of hospital discharge. One should note that the dichotomous outcome was not a rare event, with a large minority of patients dying within one year of discharge. Four thousand nine hundred and seventy (66.3%) subjects died within five years of hospital discharge.

DOI: 10.2202/1557-4679.1285

Table 1. Demographic and clinical characteristics of the 7,496 heart failure patients in the study sample.

Variable	Beta-blocker: No (N = 5,458)	Beta-blocker: Yes (N = 2,038)	P-value
	Median (25^{th} percentile – 75^{th} percentile) or N (%)	Median (25^{th} percentile – 75^{th} percentile) or N (%)	
Demographic characteristics			
Age, years	78 (70-84)	75 (67-82)	<.001
Female	2,766 (50.7%)	991 (48.6%)	0.114
Vital signs on admission			
Systolic blood pressure, mmHg	147 (127-170)	151 (130-176)	<.001
Heart rate, beats per minute	94 (78-111)	88 (73-108)	<.001
Respiratory rate, breaths per minute	24 (20-30)	24 (20-28)	<.001
Presenting signs and symptoms			
Neck vein distension	2,963 (54.3%)	1,174 (57.6%)	0.01
S3	511 (9.4%)	228 (11.2%)	0.018
S4	200 (3.7%)	87 (4.3%)	0.225
Rales > 50% of lung field	551 (10.1%)	228 (11.2%)	0.168
Findings on chest X-Ray			
Pulmonary edema	2,734 (50.1%)	1,117 (54.8%)	<.001
Cardiomegaly	1,995 (36.6%)	699 (34.3%)	0.07
Past medical history			
Diabetes	1,839 (33.7%)	788 (38.7%)	<.001
CVA/TIA	866 (15.9%)	333 (16.3%)	0.619
Previous MI	1,783 (32.7%)	966 (47.4%)	<.001
Atrial fibrillation	1,649 (30.2%)	519 (25.5%)	<.001
Peripheral vascular disease	677 (12.4%)	298 (14.6%)	0.011
Chronic obstructive pulmonary disease	1,067 (19.5%)	186 (9.1%)	<.001
Dementia	421 (7.7%)	90 (4.4%)	<.001
Cirrhosis	46 (0.8%)	6 (0.3%)	0.011
Cancer	652 (11.9%)	192 (9.4%)	0.002
Electrocardiogram – first available within 48 hours			
Left bundle branch block	823 (15.1%)	288 (14.1%)	0.304

Laboratory tests			
Hemoglobin, g/L	124 (110-138)	125 (111-138)	0.159
White blood count, 10E9/L	9 (7-12)	9 (7-11)	0.25
Sodium, mmol/L	139 (136-141)	139 (137-141)	0.002
Potassium, mmol/L	4 (4-5)	4 (4-5)	0.138
Glucose, mmol/L	7 (6-10)	8 (6-12)	<.001
Blood urea nitrogen, mmol/L	8 (6-12)	8 (6-12)	0.581
Creatinine, μmol/L	104 (82-142)	106 (85-143)	0.002

4. Traditional measures of effect in observational studies

The absence of randomization in observational studies frequently results in the baseline characteristics of treated subjects differing systematically from those of untreated subjects. If the distribution of prognostically important baseline covariates differs between treated and untreated subjects, then treatment assignment is said to be confounded with baseline covariates. As a result, direct comparisons of outcomes between treated and untreated subjects may result in biased estimates of the effect of treatment on outcomes. Health researchers have frequently used regression analysis to adjust for systematic differences between treated and untreated subjects when estimating the effect of treatment on outcomes. In this section, we briefly describe the use of regression adjustment for estimating treatment effects in observational studies. In particular, we highlight some of the limitations of this approach when outcomes are binary or time-to-event in nature.

4.1. Continuous outcomes

If the outcome is continuous, then the following linear regression model can be fit to the study data:

$$Y_i = \beta_0 + \beta_T T_i + \beta \mathbf{X}_i + e_i \quad (1)$$

where Y_i denotes the subject-specific outcome, T_i is an indicator variable denoting treatment status (T = 1 treated; T = 0 untreated), and $\mathbf{X}_i$ is a vector denoting measured baseline covariates. The regression coefficient β_T denotes the adjusted difference in the mean outcome between treated and untreated subjects. Conditional on a given value of the vector of baseline covariates, the mean of the distribution of the outcome is β_T units greater amongst treated subjects than it is in untreated subjects. Thus, the measure of treatment effect is in the same metric

DOI: 10.2202/1557-4679.1285

as that produced in an RCT with the same outcome: a difference in means. Furthermore, due to the collapsibility of the mean, the adjusted difference in means coincides with the marginal or population-average difference in means (Greenland, 1987). This can be seen because if a given subject were untreated, the expected outcome would be: $\beta_0 + \beta \mathbf{X}_i$, whereas if the same subject were treated, the expected outcome would be: $y_i = \beta_0 + \beta_T + \beta \mathbf{X}_i$. If one integrates the first quantity over the distribution of X, one obtains that the average response in the population, if all subjects were untreated, would be: $\beta_0 + \beta \int_X X_i dX$. In integrating the second quantity over the distribution of X, one obtains that the average response in the population, if all subjects were treated, would be: $\beta_0 + \beta_T + \beta \int_X X_i dX$. The difference between these two marginal responses is β_T. Therefore, the marginal effect of treatment coincides with the adjusted treated effect when the outcome is continuous.

Therefore, when linear regression adjustment is used to estimate the adjusted effect of treatment on a continuous outcome, the metric of the treatment effect is identical to that which would be obtained in an RCT in a similar context (a difference in means). Furthermore, the adjusted difference in means obtained from a linear regression model coincides with the marginal effect that would be obtained in an RCT. Finally, as in RCTs, the above method can be adapted to allow for adjustment of the baseline value of the response variable, or to estimate the effect of treatment on the change from baseline. One can adjust for the baseline value of the response variable by including it as a covariate in the regression model described in formula (1). One can also estimate the effect of treatment on the change in baseline by replacing the response variable in formula (1) by the difference between the pre- and post-intervention values of the response variable.

4.2. Binary and time-to-event outcomes

In medical research, outcomes are frequently binary or time-to-event in nature. In such settings, logistic regression models or Cox proportional hazards models are frequently used to estimate an adjusted treatment effect. The resultant measure of treatment effect is the adjusted odds ratio or the adjusted hazards ratio, when outcomes are binary and time-to-event, respectively. However, there are several limitations to the conventional use of regression models to estimate treatment effects in these contexts.

The first limitation to the conventional use of regression adjustment is that the resultant measure of treatment effect (an adjusted odds ratio or an adjusted

hazard ratio) is a relative measure of treatment effect. As such, it does not provide any information about the absolute risk of the outcome, nor of the absolute reduction in the risk of the event. As noted in Section 2, many clinical commentators have suggested that relative measures of treatment effect provide, at best, limited information about the effectiveness of a given treatment or exposure. Several commentators have suggested that absolute measures of treatment effect and the associated NNT are more informative for clinical decision making than are relative measures of treatment effect.

A second limitation to the use of regression adjustment in this context is that the adjusted odds ratio or hazard ratio is not collapsible (Greenland, 1987; Gail et al., 1984). Therefore, the adjusted odds ratio or hazard ratio does not coincide with the marginal (or population-average) odds ratio or hazard ratio that would be estimated in an RCT in a similar context. Gail et al. (1984) demonstrated that in an RCT with binary or time-to-event outcomes, the crude (marginal) odds ratio or hazard ratio does not coincide with an adjusted odds ratio or hazard ratio.

A third limitation to the use of logistic regression when the outcome is binary is also related to the odds ratio being the resultant measure of treatment effect. Many clinical readers are tempted to interpret the odds ratio as a relative risk. However, when the outcome is common, then the odds ratio is further from unity than is the relative risk (Localio et al., 2007). Thus, interpreting the odds ratio as a relative risk may result in an overestimation of the magnitude of the effect of treatment on the outcome.

We used the data described in Section 3 to illustrate the use of conventional logistic regression. The occurrence of the dichotomous outcome (death within 365 days of hospital discharge) was regressed on an indicator variable denoting receipt of a prescription for a beta-blocker at hospital discharge and the 28 baseline covariates described in Table 1. The logistic regression model assumed linear relationships between each of the continuous covariates and the log-odds of death. The odds ratio for the effect of beta-blocker prescribing on mortality was 0.73 (95% confidence interval: 0.64 – 0.83). Thus, assuming no unmeasured confounders, prescribing a beta-blocker at hospital discharge decreased the odds of death by 27% (all effects of treatment on the dichotomous outcome are summarized in Table 2). Similarly, a Cox proportional hazards regression model was used to estimate the effect of beta-blocker prescribing on post-discharge survival. The resultant adjusted hazards ratio was 0.78 (95% CI: 0.72 – 0.83). Thus, assuming no unmeasured confounders, beta-blocker prescribing reduced the hazard of death by 22%.

DOI: 10.2202/1557-4679.1285

Table 2. Comparison of treatment effects for binary outcome in case study

Method of estimation	Estimated treatment effect (95% confidence interval)
Odds ratio	
Logistic regression adjustment	0.73 (0.64, 0.83)
Relative risk	
Zhang-Yu substitution method	0.79 (0.72, 0.87)
Poisson regression	0.81 (0.73, 0.90)
Poisson regression – sandwich variance estimation	0.81 (0.74, 0.89)
Conditional standardization by centering covariates	0.77 (0.68, 0.86)
Marginal probabilities	0.81 (0.75, 0.88)
Propensity score matching	0.79 (0.71, 0.89)
Propensity score stratification	0.80 (0.73, 0.89)
Inverse probability of treatment weighting	0.81 (0.73, 0.91)
Risk difference	
Bender and Blettner	-0.063 (-0.087, -0.040)
Marginal probabilities	-0.054 (-0.076, -0.034)
Imbens	-0.053 (-0.076, -0.029)
Propensity score matching	-0.055 (-0.082, -0.029)
Propensity score stratification	-0.057 (-0.081, -0.033)
Inverse probability of treatment weighting	-0.054 (-0.077, -0.031)
Number needed to treat	
Bender and Blettner	16.0 (11.5, 26.4)
Marginal probabilities	18.5 (13.2, 29.4)
Imbens	18.9 (13.2, 34.5)
Propensity score matching	18.2 (12.2, 34.5)
Propensity score stratification	17.5 (12.3, 30.3)
Inverse probability of treatment weighting	18.5 (13.0, 32.3)

5. Regression-based methods to estimate clinically-meaningful measures of treatment effect in observational studies

Given the limited ability of direct regression adjustment to estimate clinically-meaningful measures of treatment effect when outcomes are binary or time-to-event in nature, we now review alternate regression-based methods to estimate clinically-meaningful measures of treatment effect. In this section, our focus is on settings in which outcomes are binary or time-to-event in nature.

5.1 Binary outcomes

Several authors have described methods to estimate relative risks, risk differences (or absolute risk reductions) and numbers needed to treat (NNT) using regression-based methods. We describe methods for relative risks and absolute risk reductions separately. The NNT can be estimated directly from the absolute risk reduction.

5.1.1 Relative risks

Zhang and Yu (1998) described a method to estimate an adjusted relative risk from an adjusted odds ratio that was obtained from a logistic regression model in a cohort study. Let OR denote the odds ratio obtained from a logistic regression model in which the dichotomous outcome was regressed on an indicator variable denoting treatment status and a set of covariates. Furthermore, let P_0 denote the proportion of untreated subjects who experience the outcome of interest. Zhang and Yu proposed following estimate of the relative risk:

$$RR_{\text{Zhang and Yu}} = \frac{OR}{(1-P_0)+(P_0 \times OR)}$$

They suggested that confidence intervals for the relative risk could be obtained by substituting the endpoints of the confidence interval for the adjusted odds ratio in the above formula. The approach of Zhang and Yu has been criticized by McNutt et al. (2003), as it results in estimated relative risks that are biased away from the null, with the estimated association appearing to be stronger than is true. McNutt et al. suggest that the bias in the method of Zhang and Yu arises from using one summary value of the probability of the outcome (P_0), rather than accounting for the more complex relationship between the occurrence of the outcome and treatment for each covariate pattern. Furthermore, the method of producing confidence intervals proposed by Zhang and Yu has been criticized for resulting in intervals that are artificially narrow (Localio et al., 2007; McNutt et al., 2003). Using the estimated adjusted odds ratio of 0.73 that was obtained above, and with the probability of the outcome amongst untreated subjects being 0.30, then $RR_{\text{Zhang and Yu}}$ = 0.79 (95% CI: 0.72 – 0.87). As noted above, the estimated treatment effect for binary outcomes are summarized in Table 2.

To address the limitation of the method of Zhang and Yu, McNutt et al. (2003) proposed three methods for calculating the relative risk in prospective observational studies: stratification, using a log-binomial generalized linear model, or using Poisson regression. Stratification involves estimating stratum-specific estimates of the relative risk and then pooling these stratum-specific relative risks. While this approach is attractive when there are only a few categorical confounders, it is not practical in settings in which there are many

DOI: 10.2202/1557-4679.1285

confounding covariates, some of which may be continuous in nature. Log-binomial models are a generalized linear model with a logarithmic link function and a Binomial distribution for outcomes. However, these models can be difficult to implement in practice, since the logarithmic link function only restricts the probability of an outcome to be greater than 0 (while the logit link function in a logistic regression model restricts the probability of an outcome to lie between 0 and 1). Therefore, estimation methods for these models may fail in specific empirical settings. McNutt et al. suggest that Poisson regression be used to estimate relative risks; however, they state that this approach may result in confidence intervals that are conservative when outcomes are common. In order to address the conservative nature of confidence intervals obtained using Poisson regression, Zou (2004) suggested using Poisson regression models with robust variance estimates to obtain estimates of relative risks and the associated confidence intervals. In a limited set of simulations, this approach was found to perform satisfactorily (Zou, 2004).

Using our sample dataset, we estimated the reduction in the probability of death within one year associated with beta-blocker use. Stratification was not feasible given that we wanted to account for 28 baseline covariates, several of which were continuous. The log-binomial model did not converge. The conventional Poisson model resulted in a rate ratio of 0.81 (95% CI: 0.73 – 0.90). When robust variance estimates were obtained, the resultant 95% confidence interval changed to: 0.74 – 0.89.

Wacholder (1986) suggested using a log-binomial generalized linear model estimated using a modified algorithm to estimate relative risks in prospective studies. As discussed in the above paragraph, a limitation to the use of the log-binomial model is that predicted probabilities are not constrained to lie within the unit interval. Wacholder addressed this limitation by developing a modification of the iterative estimation procedure in which the parameter space is restricted to those values that result in predicted probabilities lying within the unit interval. However, Wacholder's method requires using macros for the GLIM software, which is no longer commercially available. For this reason we did not examine estimation using this method in our sample.

Localio et al. (2007) proposed an approach to estimating the relative risk from a logistic regression model. Assume that the following logistic regression model was fit to the data:

$$\log\left(\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right) = \alpha_0 + \beta T_i + \boldsymbol{\alpha}\mathbf{X_i} \qquad (2)$$

If the baseline covariates are set to their reference values ($\mathbf{X_i} = \mathbf{0}$), then we have that $\log\left(\frac{\Pr(Y_i=1)}{1-\Pr(Y_i=1)}\right) = \alpha_0 + \beta$ for treated subjects and $\log\left(\frac{\Pr(Y_i=1)}{1-\Pr(Y_i=1)}\right) = \alpha_0$ for untreated subjects. Thus, $\Pr(Y_i=1) = \exp(\alpha_0+\beta)/(1+\exp(\alpha_0+\beta))$ for treated subjects, while $\Pr(Y_i=1) = \exp(\alpha_0)/(1+\exp(\alpha_0))$ for untreated subjects. The ratio of these two probabilities is the relative risk: $RR = (1+\exp(-\alpha_0))/(1+\exp(-\alpha_0-\beta))$. A limitation to this approach is that one can only compute the relative risk for the reference covariate pattern. Computing the relative risk for a different covariate pattern requires that the covariates be transformed so that the desired covariate pattern is the reference pattern. Localio et al. refer to this method as conditional standardization by centering covariates. A consequence of this method is that there is no longer a uniform relative risk, but potentially a different relative risk for each covariate pattern. We used this method in our sample. The reference subject was taken to be a subject whose continuous covariates were set equal to the sample mean, while the dichotomous covariates were set equal to the sample mode. The resultant relative risk was 0.77 (95% CI: 0.68 – 0.86). The 95% confidence interval was estimated using non-parametric bootstrap methods with 1,000 bootstrap replicates, as this was found by Localio et al. to have superior performance compared to competing approaches.

Finally, both Austin (2010b) and Localio et al. (2007) suggested a method based upon determining marginal probabilities of the outcome using predicted probabilities derived from a logistic regression model. Localio et al. refer to this approach as marginal standardization using logistic regression, while Austin refers to it as determining marginal probabilities from a logistic regression model. We summarize the proposed approach as follows. Assume that the logistic regression model described in formula (2) was fit to the data. Using the fitted logistic regression model, one can determine the probability of the outcome if a given subject were treated and if the same subject were untreated. The probability of the outcome if a subject were treated is

$$\frac{e^{\alpha_0+\beta+\boldsymbol{\alpha}\mathbf{X}_i}}{1+e^{\alpha_0+\beta+\boldsymbol{\alpha}\mathbf{X}_i}} \tag{3}$$

The probability of the outcome if a subject were not treated is

$$\frac{e^{\alpha_0+\boldsymbol{\alpha}\mathbf{X}_i}}{1+e^{\alpha_0+\boldsymbol{\alpha}\mathbf{X}_i}} \tag{4}$$

DOI: 10.2202/1557-4679.1285

One then determines the mean probability of the outcome if the whole population were treated and again if the whole population were untreated. These mean probabilities have been referred to as the marginal probabilities of success for treated and untreated subjects, respectively. They are marginal probabilities, since they describe the average risk in the population if the entire population were either treated or untreated. Let $\overline{p}_{T=1}$ and $\overline{p}_{T=0}$ denote the marginal probabilities of success for treated and untreated subjects, respectively. Then the relative risk can be estimated as $\frac{\overline{p}_{T=1}}{\overline{p}_{T=0}}$, while the relative risk reduction is defined as $100 \times \frac{\overline{p}_{T=0} - \overline{p}_{T=1}}{\overline{p}_{T=0}}$. Both Austin and Localio et al. proposed using bootstrap methods to estimate confidence intervals for the relative risk. We applied this method to our study sample. We used a logistic regression in which one year mortality was regressed on an indicator variable denoting treatment and the 28 covariates listed in Table 1. Non-parametric bootstrap methods with 1,000 bootstrap replicates were used to derive 95% confidence intervals. The resultant relative risk was 0.81 (95% CI: 0.75 – 0.88).

5.1.2 Absolute risk reductions and number needed to treat

Four different regression-based approaches have been suggested in the literature for estimating absolute risk reductions in cohort studies. The first was by Wacholder (1986), who suggested using generalized linear models with the identify link function to estimate risk differences. Using generalized linear models with a Bernoulli distribution and an identify link function allows one to estimate a risk difference. Limitations to this approach are that the identify link function does not constrain the predicted probabilities of an outcome to lie between 0 and 1. Instead, the predicted probabilities are allowed to take any value on the real line. In practice, this can lead to computational problems. Wacholder proposed a modification to the standard iterative estimation procedure to restrict the coefficient space to those coefficients that resulted in fitted probabilities in the unit interval. However, implementation of this method requires using macros in the GLIM software programme. For this reason, we did not consider this method further.

The second approach to estimating NNTs has been advocated by Bender and Blettner (2002) (Bender and Blettner appear to have coined the term *number needed to be exposed* for use in epidemiological applications). A logistic regression model relating the odds of the outcome to an indicator variable denoting treatment/exposure and baseline covariates is fit to the study data. When

the adjusted odds ratio is greater than one (i.e. exposure increases the odds of the outcome), then the number needed to be exposed for harm (NNEH) is obtained by combining the adjusted odds ratio (OR) from the estimated logistic regression model and the event rate in the unexposed subjects (UER):

$$\text{NNEH} = \frac{1}{(\text{OR} - 1) \times \text{UER}} + \frac{\text{OR}}{(\text{OR} - 1) \times (1 - \text{UER})} \tag{5}$$

The UER is determined as the mean probability of the outcome in untreated subjects as derived from the logistic regression model. If the adjusted odds ratio is less than one, (i.e. exposure decreases the odds of the outcome) then the number needed to be exposed for benefit (NNEB) is given by NNEB = -NNEH. Bender et al. proposed using the multivariate Delta approach to estimate confidence intervals for the adjusted NNEH (Bender and Blettner, 2002; Bender and Kuss, 2003). When applied to our sample, the NNEB was 16.0 (95% CI: 11.5 – 26.4). Thus, 16 patients would have to be treated with beta-blockers at discharge to avoid one death within one year of discharge. The risk difference can be obtained by taking the reciprocal of the NNT: the absolute risk reduction was -0.063 (95% CI: -0.087 to -0.040).

The third approach that has been advocated by different authors is based on the use of marginal probabilities, as described in Section 5.1.1. This approach has been advocated by both Bender and colleagues (2007) and by Austin (2010b). Using the terminology of the previous section, the risk difference can be estimated as $\overline{p}_{T=0} - \overline{p}_{T=1}$. There are minor variations between the approaches suggested by the different authors. The approach described above reflects the method proposed by Austin (2010b). Austin suggested averaging the predicted probabilities across the entire sample, allowing one to estimate an average treatment effect: this is the average effect at the population level of moving the entire population from untreated to treated. Bender and colleagues (2007) suggest averaging the predicted probabilities over either the treated subjects or the untreated subjects. Averaging the predicted probabilities over the treated subjects allows one to estimate the average treatment effect for the treated. Furthermore, Austin proposed using bootstrap methods to estimate associated confidence intervals, while Bender et al. (2007) proposed using the multivariate Delta method to estimate the variance of the estimated treatment effect. In our sample data, we use the method proposed by Austin, in which averaging was done over the entire sample. Ninety-five percent confidence intervals were obtained using non-parametric bootstrap methods with 1,000 bootstrap replicates. The resultant risk difference was -0.054 (95% CI: -0.076 to -0.034). Thus, treatment at discharge with a beta-blocker prescription reduced the absolute risk of death within one year by 5.4%. The associated NNT was 18.5 (95% CI: 13.2 – 29.4).

DOI: 10.2202/1557-4679.1285

Imbens (2004) suggested an approach, which, while not explicitly intended to estimate risk differences, can be easily adapted for this approach. The approach proposed by Imbens is similar to that of Austin (2010b) and Bender et al (2007). The primary difference is that rather than fit a single logistic regression model relating outcomes to treatment and baseline covariates, two separate logistic regression models are fit. Each model relates outcomes to baseline covariates. The first model is fit using only the untreated subjects, while the second model is fit using only the treated subjects. Estimates of the predicted probability of the outcome are then obtained for each subject in the sample using the first model and the second model. These will be the probabilities of the outcome if each subject were untreated and treated, respectively. These quantities can then be averaged, either over the entire sample or over the subjects who were treated. From these marginal probabilities, one can compute the absolute risk reduction. Using this method on our study sample, resulted in a risk difference of -0.053 (95% CI: -0.076 to -0.029) (1,000 bootstrap replicates were used to estimate the sampling variance of the estimated risk difference). Thus, the associated NNT was 18.9 (95% CI: 13.2 – 34.5).

5.2 Time-to-event outcomes

In this sub-section, we focus on using regression-based approaches to estimate absolute measures of treatment effect for time-to-event or survival outcomes.

5.2.1 Absolute risk reduction and number needed to treat

Austin has proposed a method to estimate an absolute risk reduction and the associated NNT from an adjusted survival model (Austin, 2010c). The method is similar to the method for estimating marginal probabilities using a logistic regression model that was described in Sections 5.1.1 and 5.1.2. The method is applicable with either the Cox proportional hazards regression model or with parametric accelerated failure time (AFT) models. Using this approach, survival time is regressed on treatment status and baseline covariates. For each subject, a predicted survival curve is generated assuming that the subject was untreated. These survival curves are then averaged, resulting in a marginal survival curve. This marginal survival curve describes survival in the population if the entire population was untreated. Then, a second survival curve is generated assuming that each subject was treated. These survival curves are then averaged, resulting in a second marginal survival curve. This marginal survival curve describes survival in the population if the entire population was treated. This approach is similar to the corrected group prognosis method for computing adjusted survival curves that was proposed by Ghali et al. (2001). For any duration of follow-up, one can

estimate the absolute probability of the event occurring in the population if the entire population was untreated and again if the entire population was treated. The difference between these two marginal probabilities is the absolute reduction in the probability of the event occurring within a specified duration of follow-up. The number needed to treat to avoid one event occurring within the specified duration of follow-up is the reciprocal of this quantity.

We used a Cox proportional hazards regression model to regress survival time on an indicator variable denoting treatment status and the 28 baseline characteristics described in Table 1. We estimated the difference in the probability of survival to 1,025 days (the median duration of follow-up in the study sample). The sampling variability of the risk difference was determined using 500 bootstrap samples. The resultant risk difference was -0.077 (95% CI: -0.097 to -0.057). Thus, one would need to prescribe a beta-blocker to 13 (95% CI: 10.3 to 17.5) patients at discharge to avoid one death within 1,025 days of discharge.

5.2.2 Expected (mean) survival time

An approach similar to that described in Section 5.2.1 can be used to estimate the effect of treatment on mean survival time. As above, the two marginal survival curves can be estimated: the survival curve if the entire population were untreated and the survival curve if the entire population were treated. The area under each marginal survival curve is the expected survival: the mean survival time in the population if all subjects were either untreated or treated. The difference between these two expected survival times is the change in expected survival time due to treatment. We do not illustrate this method in our sample data as the duration of follow-up (five years) was inadequate to reliably estimate expected survival time.

6. Propensity-score based methods to estimate clinically-meaningful measures of treatment effect in observational studies

The previous section described regression-based approaches to estimating risk differences and relative risks. In this section, we describe approaches based on the propensity score.

6.1 Definitions and background

Given an observational study with a variable T denoting treatment assignment (T =1 treated; T = 0 untreated) and a vector **X** of observed baseline covariates, the propensity score is defined to be the probability of treatment assignment conditional on the observed baseline covariates: $Z = \Pr(T = 1 \mid \mathbf{X})$ (Rosenbaum

DOI: 10.2202/1557-4679.1285

and Rubin, 1983, 1984). The propensity score is a balancing score: conditional on the propensity score, the distribution of observed baseline covariates is independent of treatment assignment (Rosenbaum and Rubin, 1983). Therefore, the distribution of observed baseline covariates will be the same between treated and untreated subjects with the same propensity score. The propensity score can be used in four different ways to estimate treatment effects: propensity-score matching, stratification on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score (Rosenbaum and Rubin, 1983; Rosenbaum 1987; Austin and Mamdani, 2006). The first three of these allow one to directly estimate clinically-meaningful measures of treatment effect.

6.2 Propensity score methods

In this section, we briefly describe propensity-score matching, stratification on the propensity score, and IPTW using the propensity score.

In propensity-score matching, pairs of treated and untreated subjects are formed such that matched subjects have similar values of the propensity score (alternatives to pair-matching include many-to-one matching on the propensity score and full matching). Estimates of treatment effect are estimated in the resultant matched sample. Propensity-score matching is frequently used in the medical literature (Austin, 2007a, 2008a, 2008b). Stratification on the propensity score stratifies the original sample according to the values of the propensity score. A commonly-used approach is to stratify the sample according to the quintiles of the propensity score (Rosenbaum and Rubin, 1984). Using this approach, five approximately equally sized strata are formed. Within each stratum, the effect of treatment on outcomes is estimated by comparing outcomes between treated and untreated subjects within that stratum. Stratum-specific estimates of treatment effect are then pooled to obtain an overall measure of treatment effect (Rosenbaum and Rubin, 1984). In propensity score weighting, the inverse probability of treatment is defined to be: $w = \frac{T}{Z} + \frac{1-T}{1-Z}$, where Z denotes the estimated propensity score. The sample is then weighted by the inverse probability of treatment (Lunceford and Davidian, 2004). In the sample weighted by the inverse probability of treatment, treatment assignment is independent of measured baseline covariates.

Propensity score methods allow one to separate the design of an observational study from the analysis of an observational study (Rubin, 2007). By matching on the propensity score, stratifying on the propensity score, or weighting by the inverse probability of treatment, the confounding between treatment status and observed baseline covariates has been eliminated. Therefore,

on average, there is no need to adjust for differences in baseline covariates between treated and untreated subjects. Rather, outcomes can be directly compared between treated and untreated subjects, as one would do in an RCT. Thus outcomes can be directly compared between treated and untreated subjects in the propensity-score matched sample, the inverse probability of treatment weighted sample, and within strata defined by the propensity score.

6.3 Using the propensity score to estimate clinically-meaningful measures of treatment effect

6.3.1 Binary outcomes

In propensity-score matching, outcomes can be directly compared between treated and untreated subjects in the matched sample. When outcomes are binary, risk differences and relative risks can be estimated directly by comparing the estimated probability of the outcome between treated and untreated subjects in the matched sample. However, one must account for the matched nature of the sample when estimating the variance of the treatment effect (Austin, 2009b).

In stratification on the propensity score, the outcomes between treated and untreated subjects can be directly compared within each propensity score stratum. Stratum-specific risk differences can then be pooled across the strata (Rosenbaum and Rubin, 1984). For relative risks, there are two options. First, stratum-specific relative risks can be formally pooled using the Mantel-Haenszel estimator of the pooled relative risk (Breslow and Day, 1987; Austin, 2008c). Second, the probability of the outcome occurring in untreated subjects can be averaged across propensity score strata. Similarly, the probability of the outcome occurring in treated subjects can be averaged across propensity score strata. These are the marginal probabilities of the outcome occurring in untreated and treated subjects respectively. The ratio of these marginal probabilities can be determined, with confidence intervals estimated using bootstrap methods. Estimates from these two stratification-based approaches should coincide if subject-specific relative risks are uniform. Lunceford and Davidian (2004) examine stratification estimators for estimating treatment effects when estimating treatment effects for continuous outcomes. Many of the estimators can be simply modified for estimating absolute risk reductions when outcomes are binary.

When using IPTW using the propensity score, outcomes can be compared directly between treated and untreated subjects in the weighted sample since the confounding between treatment and observed baseline covariates has been removed by weighting. Lunceford and Davidian examine a variety of weighting estimators for the effect of treatment on continuous outcomes. These can be modified to estimate the effect of treatment on binary outcomes.

DOI: 10.2202/1557-4679.1285

In our study sample, the propensity score was estimated using logistic regression to regress an indicator variable denoting treatment assignment on the 28 covariates listed in Table 1. The logistic regression model assumed that all subjects in the sample were independent of one another. There is no clear consensus in the statistical literature as to how best to account for hierarchically-structured data when estimating the propensity score.

An important component of any propensity score analysis is to assess the comparability of measured baseline covariates between treated and untreated subjects with a similar propensity score (Austin, 2009c). Assessment of baseline balance for this sample and this propensity score model has been reported elsewhere (Austin, 2009a). When using propensity score matching, subjects were matched on the logit of the propensity score using calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score (Austin, 2010d). Two thousand and twenty-five (99.4%) of the 2,039 treated subjects were successfully matched to an untreated subject. In the matched sample, the probability of death within one year of discharge was 0.269 and 0.214 in untreated and treated patients, respectively. In the matched sample the risk difference due to treatment was -0.055 (95% CI: -0.082 to -0.029). The associated NNT was 18.2 (95% CI: 12.2 to 34.5). The relative risk was 0.79 (95% CI: 0.71 – 0.89). Thus, treatment reduced the risk of death within one year of discharge by 21%. We then stratified the full sample on the quintiles of the propensity score. The pooled estimate of the risk difference was -0.057 (95% CI: -0.081 to -0.033). The associated NNT was 17.5 (95% CI: 12.3 to 30.3). The estimate of the relative risk obtained by estimating marginal probabilities was 0.80 (95% CI: 0.73 to 0.89). When using IPTW, the estimated risk difference was -0.054 (95% CI: -0.077 to -0.031). The associated NNT was 18.5 (95% CI: 13.0 to 32.3). Finally, using IPTW, the estimated relative risk was 0.81 (95% CI: 0.73 to 0.91).

6.3.2 Time-to-event outcomes

When using propensity-score matching and outcomes are time-to-event in nature, Kaplan-Meier survival curves can be produced separately, for treated and untreated subjects in the matched sample. The log-rank test should not be used for testing whether the two curves are statistically significantly different from one another, as this test assumes that the two samples are independent of one another (Harrington, 2005). Instead, one can use the stratified log-rank test for comparing Kaplan-Meier curves from matched samples (Klein and Moeschberger, 1997). Comparing Kaplan-Meier survival curves between treated and untreated subjects allows one to estimate absolute risk differences for specific durations of follow-up time. Furthermore, the area under each of the two Kaplan-Meier curves can be determined, allowing one to estimate the mean survival time in each treatment

group. The difference between these two quantities is the change in mean survival due to treatment. Confidence intervals for differences in mean survival and absolute risk differences for specific durations of follow-up are likely best estimated using bootstrap methods.

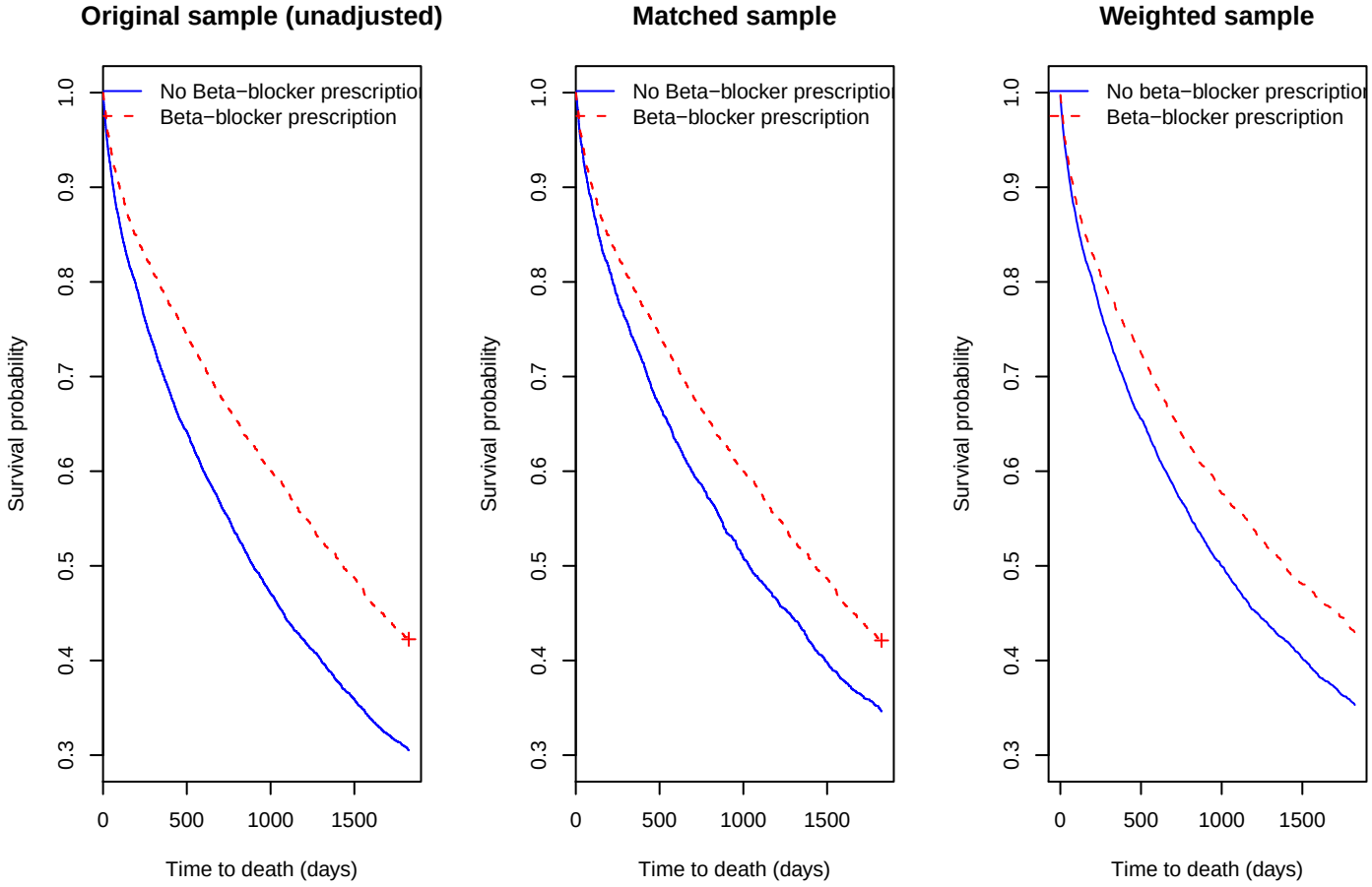
When using stratification on the propensity score, Kaplan-Meier curves can be estimated separately for treated and untreated subjects within each propensity-score stratum. Stratum-specific estimates of the absolute reduction in the risk of an event occurring within a specified duration of follow-up can then be estimated. These stratum-specific estimates can be pooled across strata. Similarly, stratum-specific estimates of the change in expected survival time can be determined. These estimates can be pooled across strata to obtain an overall estimate of the effect of treatment on expected survival time.

Xie and Liu (2005) describe an estimator of the Kaplan-Meier survival curve (and associated log-rank test) for use with inverse probability of treatment weighting. By estimating Kaplan-Meier curves in treated and untreated subjects separately, one can estimate the absolute reduction in the risk of an event occurring within a specified duration of follow-up time. Similarly, by estimating the area under each of the two survival curves, one can estimate the effect of treatment on the expected survival time.

The centre panel of Figure 1 displays the Kaplan-Meier curves comparing 5-year survival in the two treatment groups in the propensity-score matched sample. The two survival curves were statistically significantly different from one another ($P < 0.0001$). The right panel of Figure 1 displays the adjusted Kaplan-Meier curves in the weighted sample. Using the adjusted log-rank test for use with IPTW, the two curves were statistically significantly different from one another ($P < 0.0001$). For comparative purposes, the left panel of Figure 1 displays the crude Kaplan-Meier survival curves in the original study sample. In comparing the survival curves in the center and right panels to those in the left panel, one observes that the difference between the crude survival curves in the two treatment groups was reduced when confounding was accounted for.

DOI: 10.2202/1557-4679.1285

Figure 1. Kaplan−Meier survival curves in original, matched and weighted samples



7. Discussion

Randomized controlled trials (RCTs) are the gold standard for estimating the effect of treatments, interventions, and exposures on health outcomes. In RCTs, there will, on average, be no systematic differences in baseline characteristics between treated and untreated subjects. This allows outcomes to be compared directly between the different treatment arms, permitting the reporting of clinically-meaningful measures of treatment effect. In particular, absolute measures of treatment effect can be estimated when outcomes are binary and time-to-event in nature. Several clinical commentators have suggested that absolute measures of treatment effect are superior to relative measures of treatment effect for making treated-related decisions for patients (Laupacis et al.,1988; Cook and Sackett, 1995; Jaeschke et al., 1995), while others have criticized the reporting of odds ratios in RCTs (Sackett et al., 1996). At the very least, the reporting of relative measures of treatment effect should be supplemented by the reporting of absolute measures of treatment effect (Schechtman, 2002; Sinclair and Bracken, 1994).

There is growing interest in using observational or non-randomized studies to examine the effect of treatment on health outcomes. However, in non-randomized studies, there are often systematic differences between treated and untreated subjects. Historically, researchers have used regression methods to adjust for observed systematic differences between treated and untreated subjects. A limitation to this approach is that, when outcomes are binary or time-to-event in nature, the resultant measure of treatment effect is a relative measure such as an odds ratio or a hazard ratio. We suggest that absolute measures of treatment effect should also reported for observational studies of the effect of treatment on outcomes.

In this paper we have summarized methods that have been proposed in the literature for estimating clinically-meaningful measures of treatment effect in observational studies. When outcomes are binary, we have described methods for estimating relative risks, absolute risk reductions, and the number needed to treat (NNT). When outcomes are time-to-event in nature, we have described methods for estimating the absolute reduction in the probability of an event occurring within a specified duration of follow-up time (and the associated NNT). We have also suggested methods for estimating the effect of treatment on expected survival time. Application of these methods allows for supplementing the reporting of relative measures of treatment effect by absolute measures of treatment effect. Furthermore, when outcomes are dichotomous, the described methods allow for

DOI: 10.2202/1557-4679.1285

the reporting of relative risks, and free one from using the odds ratio as a measure of association and effect. When outcomes are common, the odds ratio does not provide an approximation of the relative risk; rather, it magnifies the apparent association between treatment and outcome.

We have considered two different families of approaches for estimating clinically-meaningful measures of treatment effect in observational studies: regression-based approaches and propensity score-based approaches. Each clinically-meaningful measure of treatment effect can be estimated using either approach. We would argue that there are advantages to the propensity score-based approaches compared to the regression-based approaches. First, propensity-score methods reflect a design-based approach to removing confounding, whereas regression methods reflect an analysis-based approach to removing confounding (Rubin, 2007). As Rubin (2007) has argued, the use of propensity score methods allows one to separate the design of an observational study from the analysis of an observational study. Using propensity score methods, no reference is made to the outcome until the propensity score model has been specified and adequate balance in baseline covariates has been observed between treated and untreated subjects with similar propensity scores. A second advantage to propensity-score based approaches is that one can explicitly assess the degree to which confounding has been removed. When matching, stratifying or weighting using the propensity score, one can examine the similarity of treated and untreated subjects within the matched sample, within propensity-score strata or within the weighted sample, respectively (Austin, 2009c). These balance diagnostics serve as an empirical test of whether the propensity score model has been adequately specified. When using regression-based approaches it is more difficult to assess whether the outcomes model has been adequately specified, and whether confounding between treatment and baseline covariates has been removed.

We have described how three different propensity score methods can be used to estimate clinically-meaningful measures of treatment effect: propensity score matching, stratification on the propensity score, and inverse probability of treatment weighting (IPTW) using the propensity score. There are subtle differences between these methods. First, propensity-score matching allows one to estimate average treatment effects for the treated (ATT), whereas stratification and weighting allow one to estimate average treatment effects (ATE) (Imbens, 2004). However, we would note that use of different weights allows one to estimate either the ATT or the average treatment effect for the controls (ATC) when using IPTW. Furthermore, the stratification estimator can be modified to estimate the ATT (Imbens, 2004). Second, empirical studies have shown that matching and weighting eliminates a greater degree of the observed differences between treated and untreated subjects than does stratification (Austin and Mamdani, 2006; Austin et al., 2007; Austin, 2009a). Simulations have shown that

in some settings matching and weighting remove equivalent amounts of imbalance between treated and untreated subjects, while in other settings matching removes modestly more imbalance (Austin, 2009a).

In Table 2 we summarize the different estimated measures of effect for the impact of beta-blocker prescribing on death within one year of discharge in our study sample. Several observations can be made from an examination of this table. First, the adjusted odds ratio (0.73) is further from unity than are all the estimated adjusted relative risks (relative risks range from 0.77 to 0.81). This highlights the fact that the odds ratio overestimates the magnitude of the relative risk when the outcome is common. Second, apart from conditional standardization by centering covariates (which estimates the relative risk for a specific covariate pattern), the other relative risks were lay between 0.79 and 0.81. Third, when estimating risk differences, then five of the six methods resulted in qualitatively similar estimates (-0.053 to -0.057).

We have noted above, that randomization will ensure that, on average, treated and untreated subjects do not differ systematically from one another. However, in any given randomization, it is possible that residual differences may exist between treatment groups. Several authors have suggested that regression adjustment be used to adjust for potential differences in baseline covariates that are predictive of the outcome (Senn,1989; Senn, 1994; Altman and Dore, 1991; Lavori et al., 1983). When outcomes are binary or time-to-event in nature, regression adjustment results in the odds ratio or the hazard ratio being reported as the measure of treatment effect. Several of the methods described in the current paper can be directly applied to RCTs to estimate clinically-meaningful measures of effect when regression adjustment is used and outcomes are binary or time-to-event in nature (Austin, 2010b, 2010c).

In summary, the design of RCTs allows for the reporting of simple, clinically-meaningful measures of treatment effect. The recently revised CONSORT statement on the reporting of results for RCTs recommends that, for RCTs with dichotomous outcomes, both relative and absolute measures of treatment effect be reported (Schulz et al., 2010). In observational studies of the effect of treatment or exposure on outcomes, relative measures of treatment effect, such as the odds ratio or the hazard ratio, are frequently reported. In this paper we have summarized different statistical methods that allow for estimating clinically-meaningful measures of treatment effect in observational studies. We encourage researchers to report absolute risk reductions, numbers needed to treat, and relative risks when outcomes are binary. This would allow the reporting of treatment effects in observational studies to mirror what is recommended for RCTs. When outcomes are time-to-event in nature, we encourage authors to report the absolute reduction in the risk of an event occurring within a specified duration of follow-up (along with the associated number needed to treat).

DOI: 10.2202/1557-4679.1285

References

Altman DG, Dore CJ (1991). Baseline comparisons in randomized clinical trials. *Statistics in Medicine*. **10**:797-802.

Atrial Fibrillation Investigators (1994). Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. *Archives of Internal Medicine.* **154**:1449-57.

Austin PC, Mamdani MM (2006). A Comparison of Propensity Score Methods: A Case-Study Estimating the Effectiveness of Post-AMI Statin Use. *Statistics in Medicine*. **25**:2084-2106.

Austin PC (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *Journal of Thoracic and Cardiovascular Surgery*. **134**:1128-1135.

Austin PC, Grootendorst P, Anderson GM (2007). A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study. *Statistics in Medicine*. **26**:734-753.

Austin PC (2008a). A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Statistics in Medicine*. 27:2037-2049.

Austin PC (2008b). A report card on propensity-score matching in the cardiology literature from 2004 to 2006: results of a systematic review. *Circulation: Cardiovascular Quality and Outcomes* **1**:62-67.

Austin PC (2008c). The performance of different propensity score methods for estimating relative risks. *Journal of Clinical Epidemiology*. **61**:537-545.

Austin PC (2009a). The relative ability of different propensity-score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making* **29**:661-677.

Austin PC (2009b). Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-score Matched Analyses. *The International Journal of Biostatistics*: **5**: Article 13. DOI: 10.2202/1557-4679.1146

Austin PC (2009c). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. **28**:3083-3107.

Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB (2010a). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology*. **63**:142-153. DOI:10.1016/j.jclinepi.2009.06.002.

Austin PC (2010b). Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology*. **63**:2-6. DOI: 10.1016/j.jclinepi.2008.11.004.

Austin PC (2010c). Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *Journal of Clinical Epidemiology*. **63**:46-55. DOI: 10.1016/j.jclinepi.2009.03.012.

Austin PC (2010d). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*. (In-press: pre-publication version available on journal website). DOI: 10.1002/pst.433.

Bender R, Blettner M (2002). Calculating the "number needed to be exposed" with adjustment for confounding variables in epidemiological studies. *Journal of Clinical Epidemiology* **55**:525-530.

Bender R, Kuss O (2003). Confidence intervals for adjusted NNEs: A simulation study (Letter). *Journal of Clinical Epidemiology* **56**:205-206.

Bender R, Kuss O, Hildebrandt M, Gehrmann U (2007). Estimating adjusted NNT measures in logistic regression analysis. *Statistics in Medicine*. **26**:5586-5595.

Breslow NE and Day NE (1987). *Statistical Methods in Cancer Research. Volume II – The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer.

Cook RJ, Sackett DL (1995). The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal*. **310**:452-454.

DOI: 10.2202/1557-4679.1285

Cox DR, Oakes K (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Douketis JD, Ginsberg JS, Holbrook A, Crowther M, Duku EK, Burrows RF (1997). A reevaluation of the risk for venous thromboembolism with the use of oral contraceptives and hormone replacement therapy. *Archives of Internal Medicine*. **157**:1522-30.

Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ (2001). Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 285:2864-70.

Gail MH, Wieand S, Piantadosi S (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. **7**:431-444.

Ghali WA, Quan H, Brant R, van Melle G, Norris CM, Faris PD, Galbraith PD, Knudtson for the APPROACH Investigators (2001). Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA*. **286**:1494-1497.

Greenland S (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*. **125**:761-768.

Harrington D (2005). Linear rank tests in survival analysis. In: *Encyclopedia of Biostatisics, Second edition* (editors: Armitage P and Colton T). New York, NY: John Wiley & Sons; Pages 2802-2812.

Imbens GW (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* .**86**:4-29.

Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N (1995). Basis statistics for clinicians 3: Assessing the effects of treatment: measures of association. *Canadian Medical Association Journal* **152**:351-357.

Klein JP, Moeschberger ML (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer-Verlag.

Lavori PW, Louis TA, Bailar JC III, Polansky M (1983). Designs for experiments – Parallel comparisons of treatment. *New England Journal of Medicine* **309**:1291-1298.

Laupacis A, Sackett DL, Roberts RS (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*. **318**:1728-1733.

Localio AR, Margolis DJ, Berlin JA (2007). Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology*. **60**:874-882.

Lunceford JK, Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. **23**:2937-2960.

McNutt LA, Wu C, Xue X, Hafner JP (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*. **157**:940-943.

Permutt T (1990). Testing for imbalance of covariates in controlled experiments. *Statistics in Medicine*. **9**:1455-1462.

Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**:41-55.

Rosenbaum PR, Rubin DB (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*. **79**:516-524.

Rosenbaum PR (1987). Model-based direct adjustment. *The Journal of the American Statistician* **82**:387-394.

Rubin DB (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* **26**:20-36.

Sackett DL, Deeks JJ, Altman DG (1996). Down with odds ratio! *Evidence-Based Medicine*. September/October:164-166.

Schechtman E (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat – which of these should we use? *Value in Health* **5**:431-436.

DOI: 10.2202/1557-4679.1285

Schulz KF, Altman DG, Moher D, for the CONSORT Group (2010). Consort 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**:c332. DOI: 10.1136/bmj.c332.

Senn SJ (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**:467-475.

Senn S (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine* **13**:1715-1726.

Sinclair JC, Bracken MB (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology.* **47**:881-889.

Singer DE, Albers GW, Dalen JE, Fang MC, Go AS, Halperin JL, Lip GY, Manning WJ; American College of Chest Physicians (2008). Antithrombotic therapy in atrial fibrillation: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). *Chest* **133**(6 Suppl):546S-592S.

Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM (2004). *Quality of Cardiac Care in Ontario – Phase 1. Report 1*. Toronto: Institute for Clinical Evaluative Sciences.

Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT (2009). Effectiveness of public report cards for improving the quality of cardiac care: The Enhanced Feedback For Effective Cardiac Treatment (EFFECT) study. *Journal of the American Medical Association*. **302**:2330-2337.

van Walraven C, Hart RG, Connolly S, Austin PC, Mant J, Hobbs FD, Koudstaal PJ, Petersen P, Perez-Gomez F, Knottnerus JA, Boode B, Ezekowitz MD, Singer DE (2009). Effect of age on stroke prevention therapy in patients with atrial fibrillation: the Atrial Fibrillation Investigators. *Stroke*. **40**:1410-6.

Wacholder S (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology* **123**:174-184.

Xie J, Liu C (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*. **24**:3089-3110.

Zhang J, Yu KF (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*. **280**:1690-1691.

Zou G (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*. **159**:702-706.

DOI: 10.2202/1557-4679.1285