Published in final edited form as:

Biometrics. 2013 September ; 69(3): 724–731. doi:10.1111/biom.12038.

Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect

Changyu Shen^{1,*}, Jaesik Jeong¹, Xiaochun Li¹, Peng-Shen Chen², and Alfred Buxton^{3,4}

¹Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

²The Krannert Institute of Cardiology and the Division of Cardiology, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA

³Harvard Medical School, Boston, MA 02215, USA

⁴Clinical Electrophysiology Laboratory, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

Summary

It is well recognized that the conventional summary of treatment effect by averaging across individual patients has its limitation in ignoring the heterogeneous responses to the treatment in the target population. However, there are few alternative metrics in the literature that are designed to capture such heterogeneity. We propose the concept of treatment benefit rate (TBR) and treatment harm rate (THR) that characterize both the overall treatment effect and the magnitude of heterogeneity. We discuss a method to estimate TBR and THR that easily incorporates a sensitivity analysis scheme, and illustrate the idea through analysis of a randomized trial that evaluates the Implantable Cardioverter-Defibrillator (ICD) in reducing mortality. A simulation study is presented to assess the performance of the proposed method.

Keywords

Causal inference; Heterogeneity in treatment effect; Potential outcomes; Sub-group analysis

1. Introduction

Current clinical practice largely relies on knowledge obtained from medical studies that characterize the treatment (e.g. a new medicine or device) effect in an “average” sense. In a typical randomized Phase III clinical trial, often the primary interest is the contrast of the means or proportions of the outcome between the intervention and the control arms. However, there are patients who do not benefit from an intervention even when a trial is positive with statistical evidence of non-zero difference in the means or proportions, and there are patients who benefit from an intervention even when a trial is negative. Consequently, the average treatment effect (ATE) fails to capture variation in response to a treatment due to heterogeneity at many levels among patients in the target population (Davidoff, 2009; Kent and Hayward, 2007). To study heterogeneity in treatment effect, pre-specified sub-group analysis has been a common practice for the design of a clinical trial (Wang et al., 2007). Nevertheless, the same “average” metrics are again used to summarize

*chashen@iupui.edu.

Supplementary Materials

Web Appendices referenced in Section 2.3 are available with this paper at the Biometrics website on Wiley Online Library.

the treatment effect in each sub-population, omitting the heterogeneity within each sub-population. Therefore, a fundamental issue is the lack of appropriate metrics to characterize treatment heterogeneity within a given group of patients.

We make our point using the Implantable Cardioverter-Defibrillator (ICD) as an example. The ICD is a device that detects life-threatening heart rhythm disorders such as ventricular tachycardia or fibrillation, and responds by electrical stimulation or shocks to restore normal rhythm. Several multi-center clinical trials have shown that ICDs can reduce overall mortality by 20–30% as compared with anti-arrhythmic medications or placebo in low ejection fraction populations for primary prevention (Bardy et al., 2005; Moss et al., 2002). Since the ICD is not free of other health risks, including mortality caused by ICD, and is expensive to implant and maintain (Tung, Zimetbaum and Josephson, 2008), it should ideally be implanted only if it could save lives. Obviously, it is also important NOT to implant ICD if it is going to increase the risk of mortality. As will be defined in Section 2, the relative sizes of the sub-populations who will benefit from or be harmed by ICD correspond to the treatment benefit rate (TBR) and treatment harm rate (THR), respectively. For example, the Multicenter Automatic Defibrillator Implantation Trial II (MADIT-II) was designed to evaluate the potential survival benefit of the ICD in patients with a prior myocardial infarction and a left ventricular ejection fraction of 0.30 or less (Moss et al., 2002). A total of 1232 patients were assigned in a 3 to 2 ratio to receive either an ICD or the conventional medical therapy. The 2-year mortality rates are 14% and 21% for the ICD and the conventional therapy arms, respectively, translating to one-third mortality reduction by the ICD relative to the conventional therapy. Nevertheless, only the 21% in the control arm who have died may have derived survival benefit had they received ICD, indicating that the TBR cannot be more than 21%. In other words, at least 79% of the patients did not derive survival benefit from the ICD. This is quite a contrast with the seemingly more optimistic relative reduction of overall mortality. Considering the ICD-associated cost and health complications, these numbers are clearly important for decision-making from the perspectives of individual patients, physicians and public health policy.

In this article, we propose a formal framework to define TBR and THR for a binary endpoint and an estimation procedure based on the conditional independence of the potential outcomes. Our estimation strategy also naturally incorporates a sensitivity analysis scheme to assess the impact of assumption violation. In what follows, we describe the framework and estimation method in Section 2, apply the method to MADIT-II data in Section 3, assess the estimation procedure through a simulation study in Section 4, and conclude the article with a discussion section.

2. Method

2.1 Background

A well-accepted conceptual framework to study treatment effect is the potential outcomes (Holland, 1986; Rubin, 2007). In this framework, a subject is assumed to have two potential outcome values, Y_1 (under the intervention) and Y_0 (under the control). The treatment effect of the intervention is a contrast between Y_1 and Y_0 , i.e. $Y_1 - Y_0$. However, in reality often we only observe the potential outcome under the treatment the subject actually receives. The essential idea of potential outcomes, therefore, is the acknowledgement of the outcome that could have been observed had the subject received the treatment different from what s/he actually received. This practical difficulty prevents us from directly estimating the treatment effect for any given individual. Nevertheless, one can estimate the treatment effect averaged across patients through either randomized trials or observational studies (Rubin, 2007).

2.2 Definition

Roughly speaking, TBR is the proportion of the relevant population that benefits from the intervention as compared with the control for a given endpoint. THR is the proportion that is harmed by the intervention as compared with the control based on the same endpoint. Concepts of TBR and THR under specific definitions of “benefit” have been proposed in the literature (Albert, Gadbury and Mascha, 2005; Gadbury and Iyer, 2000; Gadbury, Iyer and Albert, 2004; Gadbury, Iyer and Allison, 2001). Nevertheless, previous literature has been focused on developing inferential bounds for TBR. Here we propose a general framework for the estimation of TBR/THR. Specifically, let Y_1 and Y_0 be the potential binary outcomes under the intervention and the control, respectively, with value 1 indicating more favorable health condition and 0 otherwise. Naturally, “benefit” and “harm” can be defined as

$$\text{Benefit:}(Y_0=0, Y_1=1); \text{Harm:}(Y_0=1, Y_1=0).$$

Hence the TBR and THR can be defined as

$$\text{TBR}=\Pr(Y_0=0, Y_1=1); \text{THR}=\Pr(Y_0=1, Y_1=0).$$

Obviously, the rest of the population, $1-\text{TBR}-\text{THR}$, represent those subjects whose end-points remain the same regardless of treatment received. Straight forward algebra shows the following relationship:

$$\text{TBR} - \text{THR} = \Pr(Y_1=1) - \Pr(Y_0=1).$$

Thus, the difference between TBR and THR is the marginal absolute treatment effect. In Table 1, we summarize the MADIT-II results in the potential outcomes framework, with entries representing the proportions of patients having 4 different combinations of (Y_0, Y_1). The marginal mortality rates of 14% and 21% are the mortality rates under ICD and conventional medical therapy, respectively. The TBR and THR are labeled as two of the four cell entries. The two cells on the diagonal line represent those whose mortality status remains the same regardless of ICD.

TBR and THR convey two essential pieces of information regarding the effect of an intervention. First, they provide a summary on the overall effect of an intervention as measured by the proportion of the population that derives clinical benefit or harm. Second, they automatically characterize the heterogeneity in response to a treatment by the contrast of the proportions of those who benefit from the intervention, those who are harmed by the intervention, and the rest who are not affected. Hence, the multinomial characterization of “benefit”, “harm” and the rest allows the proportions to reveal both the mean and variation.

2.3 Estimation

It is clear from the definition of TBR and THR that they depend on the joint distribution of the potential outcomes, which cannot be directly estimated using the observed data. Methods have been proposed to establish the bounds for TBR under specific definitions of benefit for binary (Albert et al., 2005; Gadbury et al., 2004) and continuous endpoints (Gadbury and Iyer, 2000; Gadbury et al., 2001). We propose a method to directly estimate TBR and THR for binary endpoints using randomized trial data based on the conditional independence of the potential outcomes. As this assumption cannot be tested using observed data, our method also naturally incorporates a sensitivity analysis scheme to assess the impact of the violation of the assumption.

We assume that conditional on a set of relevant baseline covariates (pre-treatment) X , the two potential outcomes are independent:

$$Y_0 \perp Y_1 | X. \quad (1)$$

Assumption (1) means within each stratum of X , Y_0 does not predict Y_1 and vice versa. Let TBR_X and THR_X be the treatment benefit and treatment harm rate given X , we have

$$\begin{aligned} TBR &= E(TBR_X) = E[\Pr(Y_0=0, Y_1=1|X)] = E[\Pr(Y_0=0|X)\Pr(Y_1=1|X)] \\ THR &= E(THR_X) = E[\Pr(Y_0=1, Y_1=0|X)] = E[\Pr(Y_0=1|X)\Pr(Y_1=0|X)]. \end{aligned} \quad (2)$$

Here the expectation is with respect to the marginal distribution of X , and the last equality is due to assumption (1). In a randomized trial, let A be the treatment assignment indicator such that $A = 1$ indicates intervention and $A = 0$ indicates control. The observed outcome Y can be written as $Y = AY_1 + (1 - A)Y_0$. Due to randomization, $\Pr(Y_0 = 1|X) = \Pr(Y_0 = 1|X, A = 0) = p_0(X)$ and $\Pr(Y_1 = 1|X) = \Pr(Y_1 = 1|X, A = 1) = p_1(X)$. Equation (2) becomes

$$TBR = E[(1 - p_0(X))p_1(X)] \text{ and } THR = E[p_0(X)(1 - p_1(X))]. \quad (3)$$

Equation (3) demonstrates that under assumption (1), TBR and THR are estimable using data from a randomized trial as the definition involves only the marginal distribution of each potential outcome.

Introduction of X in equations (2) and (3) reveals two sources of variation in benefit/harm status for a given population, namely, the between-stratum and within-stratum variations. Using TBR as an example, the variance (with respect to X) of TBR_X represents the variation explained by X (between-stratum). On the other hand, for a given X , $TBR_X(1 - TBR_X)$ represents the variation unexplained. As one of the reviewers pointed out, for the purpose of inducing conditional independence, X should include *prognostic* factors that predict both potential outcomes. These prognostic factors do not necessarily increase the between-stratum variation relative to the within-stratum variation. Nevertheless, this is not an issue for our method as our primary interest is the overall TBR. See Section 5 for more details.

Sometimes the binary endpoint is defined as whether or not a negative health event occurs before a fixed time threshold t based on a time-to-event outcome W . In this case,

$$p_0(X) = \Pr(W > t | X, A=0) = 1 - F_X(t) \quad p_1(X) = \Pr(W > t | X, A=1) = 1 - G_X(t). \quad (4)$$

Here, $F_X(t)$ and $G_X(t)$ are the conditional cumulative distribution functions (CDF) of W for the control and intervention arms, respectively. Then the TBR and THR can be expressed as

$$TBR(t) = E[F_X(t)(1 - G_X(t))] \text{ and } THR(t) = E[(1 - F_X(t))G_X(t)]. \quad (5)$$

Again, $TBR(t)$ and $THR(t)$ only depend on the marginal distributions of the potential outcomes and are estimable.

We now provide details of the estimation procedure for TBR. The estimation of THR can be similarly derived. Let $(X_i, A_i, Y^{(i)})$, $i = 1, \dots, n$ be the data of n subjects from a randomized trial, where X_i is a vector of baseline covariates, A_i is the treatment indicator with value 1 meaning intervention and 0 meaning control, and $Y^{(i)}$ is the observed binary outcome under

A_i : A parametric regression model (e.g. logistic regression) can be fitted to the data to obtain estimates of $p_0(X_i)$ and $p_1(X_i)$, $p_0(X_i)$ and $p_1(X_i)$. Then a natural estimator based on (3) is

$$\widehat{\text{TBR}} = \frac{1}{n} \sum_{i=1}^n (1 - \hat{p}_0(X_i)) \hat{p}_1(X_i). \quad (6)$$

With time-to-event data that are subject to right censoring, $Y^{(i)}$ is replaced by $(T_i = \min(W_i, C_i), \delta_i = \mathbb{I}(W_i \leq C_i))$ where W_i is the event time, C_i is the censoring time and $\mathbb{I}(\cdot)$ is the indicator function that takes value 1 if the argument is logically true and 0 otherwise. Then a regression model (e.g. Cox proportional hazard model) can be used to obtain estimates of $F_X(t)$ and $G_X(t)$, $F_X(t)$ and $G_X(t)$. A natural estimator of $\text{TBR}(t)$ is

$$\widehat{\text{TBR}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{F}_X(t) (1 - \hat{G}_X(t)). \quad (7)$$

If maximum likelihood estimation and maximum partial likelihood estimation are used to obtain parameter estimates in (6) and (7), both $\widehat{\text{TBR}}$ and $\widehat{\text{TBR}}(t)$ are consistent and asymptotically normal. The consistency is straight forward to establish based on the well-known properties of maximum likelihood and partial likelihood estimators, and is omitted here. In the Supplementary Materials we prove the asymptotic normality of the two estimators.

It should be noted that the stratum specific rates, TBR_X and THR_X , are important for a personalized treatment strategy or identification of sub-population(s) with distinct treatment benefit/harm profile. Research towards these directions is beyond the scope of this paper. In Section 5, we provide some discussion along these lines.

2.4 Sensitivity Analysis

Assumption (1) cannot be directly tested using the observed data as for each subject we only observe one of the two potential outcomes. This is a similar challenge to causal inference in observational studies and inference based on data that are missing not at random (Rubin, 1976). A typical solution is to construct a sensitivity analysis scheme where one postulates various assumptions on the nature of the association between Y_1 and Y_0 (conditional on X) and examine potential bias induced. For a binary endpoint, the association between Y_1 and Y_0 can be characterized by the odds ratio:

$$\gamma_X = \left[\frac{\Pr(Y_0=1, Y_1=1)\Pr(Y_0=0, Y_1=0)}{\Pr(Y_0=1, Y_1=0)\Pr(Y_0=0, Y_1=1)} \middle| X \right].$$

The assumption of conditional independence is violated when $\gamma_X \neq 1$. It can be shown that

$$\begin{aligned} \text{TBR}_X &= \Pr(Y_0=0, Y_1=1) \\ &= C(p_0(X), p_1(X), \gamma_X) \\ &= \frac{1 - (p_0(X) - p_1(X))(1 - \gamma_X) - [(1 - p_0(X) - p_1(X))^2 - 4(\gamma_X - 1)\gamma_X p_0(X)p_1(X)]^{0.5}}{2(\gamma_X - 1)}, \gamma_X \neq 1. \end{aligned}$$

A sensitivity analysis scheme for a binary endpoint is then to calculate

$$\widehat{\text{TBR}}(\gamma) = \frac{1}{n} \sum_{i=1}^n C(\hat{p}_0(X_i), \hat{p}_1(X_i), \gamma_{X_i}), \quad (8)$$

and examine how the estimate of TBR changes as X varies over possible scenarios. Similarly,

$$\widehat{\text{TBR}}(\gamma, t) = \frac{1}{n} \sum_{i=1}^n C(1 - \hat{F}_{X_i}(t), 1 - \hat{G}_{X_i}(t), \gamma_{X_i}). \quad (9)$$

Again, if the postulated odds ratio X is correct, then the estimator in (8) and (9) are consistent and asymptotically normal. Sensitivity analysis scheme for the estimation of THR can be similarly constructed.

3. Application to the MADIT-II trial

The proposed method is applied to the MADIT-II trial to estimate the TBR and THR associated with the ICD as compared to the conventional medical therapy. The endpoint in our analysis is 2-year survival since randomization. A subject is considered to derive survival benefit from ICD if s/he will survive beyond 2 years with the ICD but will die within 2 years with the conventional therapy. Similarly, ICD is considered harmful if the subject will survive beyond 2 years with conventional therapy but will die within 2 years with the ICD. To estimate TBR and THR, we consider separate Cox proportional hazard models for the ICD and conventional medical therapy arms. As a Cox proportional hazard model has been established for total mortality in the conventional medical therapy arm based on five binary prognostic factors (Goldenberg et al., 2008), we will use the same model for both arms. The five prognostic factors are listed in Table 2. In our analysis, we excluded 60 subjects with blood urea nitrogen greater or equal to 50 mg/dl or serum creatinine greater or equal to 2.5 mg/dl as this group has very high mortality risk and is quite different from the majority of the targeted population (Goldenberg et al., 2008). Thus, we focus on the rest of the 1172 subjects that represent a general population without the “outliers”. We further excluded 104 subjects with missing values in at least one of the five risk factors. Therefore, our analysis included 1172–104=1068 subjects (425 from conventional medical therapy arm and 643 from the ICD arm). Comparing the 104 subjects with the 1068 subjects within each treatment arm in terms of survival curve leads to p -values (log-rank test) of 0.09 and 0.52 for medical therapy and ICD arms, respectively. Thus, there is no strong evidence of departure in survival probability between those with missing-covariates and those without. In addition, the 104 excluded subjects consist of less than 9% of the initial sample of 1172 subjects. Therefore, the potential bias due to the exclusion of these subjects should be minimal.

The TBR and THR estimators in this case are

$$\widehat{\text{TBR}}(2) = \frac{1}{n} \sum_{i=1}^n \hat{F}_{X_i}(2)(1 - \hat{G}_{X_i}(2)) \text{ and } \widehat{\text{THR}}(2) = \frac{1}{n} \sum_{i=1}^n (1 - \hat{F}_{X_i}(2))G_{X_i}(2). \quad (10)$$

Here F_{X_i} and $\hat{G}_{X_i}$ are calculated based on the estimates of the baseline hazard and regression coefficients of the five risk factors from the Cox models of the two arms. We also consider a sensitivity analysis where the odds ratio X between the two binary potential outcomes is postulated. To simplify the analysis, we consider $X =$ for all X . To obtain reasonable

values of the odds ratio to guide the sensitivity analysis, we identified for each control subject an intervention subject with the same profile of the five risk factors through greedy search with replacement (Jasjeet, 2011). Out of the 425 pairs of subjects, there were 99 pairs whose 2-year survival status can be ascertained. If we consider the 99 pairs of binary indicators as 99 pairs of (Y_0, Y_1) , the odds ratio is 1.67. Certainly, this will be a coarse estimate of the overall association between Y_0 and Y_1 . But this gives us some basic idea on the magnitude of odds ratio that might be reasonable. Based on this result, we consider an odds ratio of 2, 3, or 4 for all X strata. This range reflects moderate to strong positive correlation.

The results are summarized in Table 3. Note that we constructed the 95% confidence intervals based on normal approximation with standard errors calculated in two ways, one based on asymptotic standard error and one based on the empirical standard error of 500 bootstrap samples. It can be seen that the two methods agree with each other very well. As the marginal survival rates are 79% and 86% for the conventional medical therapy and ICD, respectively, TBR is bounded between 7% and 21% (Albert et al., 2005). Nevertheless, these bounds have not accounted for sampling variation. With appropriate accounting of sampling variation, the real uncertainty is higher. Our TBR estimate assuming conditional independence is 18%, close to the upper bound of 21%. Obviously, an odds ratio greater than 1 reduces the TBR estimate because the positive correlation will reduce the proportion of discordant pairs (e.g. $Y_0 = 0$ and $Y_1 = 1$). It can be seen that even with a strong correlation characterized by an odds ratio of 4, the TBR estimate is still 15%, suggesting that the true TBR is likely higher than 7%. Overall, after accounting for potential assumption violation and sampling variation, the true TBR is most likely somewhere between 11% and 22% (e.g. lowest lower confidence limit to highest upper confidence limit among $=1,2,3,4$). Similarly, the THR is most likely somewhere between 4% and 12%. Here we did not consider negative correlation between the potential outcomes for two reasons. First, it will have minimal effect as the TBR estimate assuming conditional independence is already quite close to the upper bound. Second, a negative correlation is unlikely to hold for this particular problem. There is no reason to believe that the probability of surviving 2 years or more under the ICD is higher for those who would die within 2 years under conventional medical therapy as compared to those who would survive 2 years or more under the conventional medical therapy.

When the conditional independence assumption holds, TBR_X ranges from 9.4% to 41.9% with a standard deviation of 8.8% among the 32 unique X patterns observed in the data. Similarly, THR_X ranges from 6.8% to 22.5% with a standard deviation of 4.0%. When $\theta = 4$, TBR_X ranges from 8.1% to 35.5% with a standard deviation of 7.6% and THR_X ranges from 4.8% to 14.7% with a standard deviation of 2.8%. Thus, there is substantial variation in the benefit/harm profile among different X strata, suggesting strong patient heterogeneity explained by X based on our model.

Our analysis suggests that around 11–22% of the population targeted by MADIT-II actually benefits from ICD in terms of 2-year survival and 4–12% is harmed by ICD. From the perspective of each individual patient, this piece of information undoubtedly will be critical for the decision-making process that needs to balance all relevant factors. From the perspective of public health, this information also is important to establish appropriate guideline and regulation policy to optimize the overall outcome in a cost-effective manner.

4. A simulation study

We conducted a simulation study to evaluate the performance of the estimators in the setting of Section 3. Again, the primary aim is to estimate the TBR and THR in terms of 2-year survival.

To generate similar data as the example in Section 3, we first fitted separate Weibull regression models to the conventional medical therapy and ICD arms. For a given arm, the Weibull regression assumes the following model for the survival time w

$$f(w|X) \sim \text{Weibull}(\alpha, \lambda(X)) = \frac{\alpha}{\lambda(X)} \left[\frac{w}{\lambda(X)} \right]^{\alpha-1} \exp \left[- \left(\frac{w}{\lambda(X)} \right)^\alpha \right],$$

where α is the shape parameter, and $\lambda(X)$ is the scale parameter that depends on the covariate vector X (X includes five binary covariates):

$$\lambda(X) = \exp[(1, X^T)\beta], \beta = (\beta_0, \beta_1, \dots, \beta_5)^T.$$

Simple algebra shows that this model is a proportional hazard model with

$$h(w|X) = h_0(w) \exp[-\alpha\beta_1 X_1 - \dots - \alpha\beta_5 X_5], h_0(w) = \alpha \exp(-\alpha\beta_0) w^{\alpha-1},$$

where $h(\cdot|X)$ and $h_0(\cdot)$ are the hazard function given X and baseline hazard function. We obtained the parameters estimates through maximum likelihood estimation. To generate data, we treated the empirical distribution of X and the estimated Weibull models as the true data generation mechanism and kept the ratio of intervention to control at 3:2. We assumed a normal censoring process that is independent of the survival time with parameters set to maintain the same censoring rate as what was observed in the real data. Two separate simulation schemes were considered:

- i. Conditional on X , the potential survival times under intervention and control are independent. With this assumption, we do not need any further distribution assumptions to generate data.
- ii. Conditional on X , the potential survival times are not independent. We introduced a Gamma frailty with mean 1, which induces odds ratio for two-year survival between the two arms that depends on X . Under the frailty model, the proportional hazard property still holds for the marginal distribution of the survival time. The parameter for the Gamma distribution is set to the value such that the marginal odds ratio of the two binary potential outcomes is the same as what is to be assumed during the estimation process. Note that the odds ratios for X strata induced by the Gamma frailty are different, whereas the sensitivity analysis during the estimation assumes constant odds ratio. Thus, the simulation allows us to evaluate the effect of this deviation.

For each setting reported in Table 4, we generated 1000 Monte Carlo data sets each composed of 1000 data points. To interpret Table 4, consider the second row as an example. When the true marginal odds ratio of the two binary potential outcomes is 2, the true value of TBR is 0.166. Our estimation procedure based on the Cox model and the assumption of constant odds ratio of 2 across all X strata yields a bias of 0.001 and standard error of 0.020. The average of large-sample standard error estimate is 0.020, and the 95% CI using normal approximation with large-sample standard error estimate has a coverage probability of

0.961. Overall, our estimation works quite well, even though the assumption of constant odds ratio across X strata is incorrect. Such a simplification can still capture the shift of TBR and THR when the conditional independence assumption is violated.

5. Discussion

Currently well-adopted metrics for the evaluation of the effect of an intervention only capture the aggregated treatment effect, leaving the characterization of the heterogeneity in response to a treatment largely absent from the discussion of clinical decision-making and policy-making. In this article, we propose TBR and THR as a way to characterize both the overall positive/negative effect of an intervention and the heterogeneity in response to the intervention for a binary endpoint. The conceptual framework and estimation method we proposed are generic and can be extended to both ordinal and continuous endpoints. A key working assumption of the new initiative of comparative effectiveness research is that the real world setting adds another dimension of variation in treatment response as compared with a well-controlled clinical trial. This necessitates the need for better understanding of treatment heterogeneity. Hence, informative metrics like TBR and THR are in great need to facilitate the advancement toward better treatment strategy.

We want to emphasize the distinction between *prognostic* and *predictive* factors in their relationship to TBR and THR. We say X is a prognostic factor if it is associated with both Y_0 and Y_1 . On the other hand, X is a predictive factor if it is associated with the difference between Y_0 and Y_1 . A baseline variable may be prognostic, predictive, neither prognostic nor predictive, or both. One example raised by one of the reviewers is that a variable associated with Y_1 but not Y_0 is a predictive but not prognostic factor. For the purpose of introducing conditional independence in assumption (1), we essentially are after prognostic factors, some of which might not be predictive. Therefore, the TBR_X and THR_X in this manuscript do not necessarily maximize the between-stratum variation. This presents no problem as our primary purpose is to estimate the overall TBR and THR.

The concept of TBR and THR offers direct evidence for the necessity of individualized treatment. For the ICD example, there is substantial variation in TBR_X as shown in the previous section, a strong justification for an individualized strategy for ICD implementation. Apparently, only the 11–22% who will actually benefit from ICD should receive it and a more cost-effective strategy should be considered for the rest. In particular, we should not give ICD to the 4–12% who will be harmed by the device. Although TBR_X and THR_X can be directly used to select patients most likely to benefit from ICD, their estimators tend to have high variability. A natural alternative is to group patients with similar X together into a sub-population, for which the two metrics can be more accurately estimated. The problem is on the partition of the space of predictive factors into a number of strata to maximize the between-stratum variation relative to the within-stratum variation in these metrics. This type of approach has been reported in the literature using the tool of classification and regression trees (Foster, Taylor and Ruberg, 2011; Lipkovich et al., 2011).

In an ideal situation, we divide the population into three sub-populations dominated by those who benefit, those who are harmed and those who are not affected, respectively. Nevertheless, it is unrealistic in practical problems to identify sub-populations 100% homogeneous of one type of subjects. Therefore, characterization of the heterogeneity within a sub-population is meaningful, which justifies the calculation of TBR and THR within a sub-population. In addition, sub-population specific TBR and THR offer a relevant individualized measurement of the benefit/harm of an intervention, particularly from the perspective of patients/physicians. Hypothetically, suppose for some sub-population the 2-year mortality rates are 10% and 5% under conventional medical therapy and ICD,

respectively, and TBR= 5% and THR= 0%. Therefore, ICD reduces the mortality by 50%, which, by the criterion of relative risk reduction, is very effective. Nevertheless, the 5% TBR suggests that only 5% of the patients within the sub-population actually derive survival benefit from the ICD, which is fairly low (e.g. 5% is the typical threshold for a p -value to be considered statistically significant). From the perspective of patients/physicians, TBR offers a unique and relevant aspect of the problem for decision-making.

A limitation of the proposed approach is that it relies on assumptions that cannot be directly tested by the observed data. Therefore, we recommend the sensitivity analysis as an integral aspect of the methodology to assess the impact of assumption violation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported in part by National Institutes of Health (NIH) grant R21 CA152463 and the Indiana University Health-Indiana University School of Medicine Strategic Research Initiative in cardiology. The MADIT-II data were provided to this research by Boston Scientific CRM. We would like to thank Peter Lam from Boston Scientific for constructive suggestions, and anonymous referees for very useful comments that improved the presentation of the paper.

References

- Albert JM, Gadbury GL, Mascha EJ. Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biometrical Journal*. 2005; 47:662–673. [PubMed: 16385907]
- Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*. 1982; 10:1100–1120.
- Bardy GH, Lee KL, Mark DB, Poole JE, Packer DL, Boineau R, Domanski M, Troutman C, Anderson J, Johnson G, McNulty SE, Clapp-Channing N, Davidson-Ray LD, Fraulo ES, Fishbein DP, Luceri RM, Ip JH. Amiodarone or an implantable cardioverter-defibrillator for congestive heart failure. *New England Journal of Medicine*. 2005; 352:225–237. [PubMed: 15659722]
- Davidoff F. Heterogeneity is not always noise: lessons from improvement. *The Journal of the American Medical Association*. 2009; 302:2580–2586.
- Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*. 2011; 30:2867–2880. [PubMed: 21815180]
- Gadbury GL, Iyer HK. Unit-treatment interaction and its practical consequences. *Biometrics*. 2000; 56:882–885. [PubMed: 10985231]
- Gadbury GL, Iyer HK, Albert JM. Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference*. 2004; 121:163–174.
- Gadbury GL, Iyer HK, Allison DB. Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics*. 2001; 11:313–333. [PubMed: 12018782]
- Goldenberg I, Vyas AK, Hall WJ, Moss AJ, Wang H, He H, Zareba W, McNitt S, Andrews ML. Risk stratification for primary implantation of a cardioverter-defibrillator in patients with ischemic left ventricular dysfunction. *Journal of the American College of Cardiology*. 2008; 51:288–296. [PubMed: 18206738]
- Holland PW. Statistics and causal inference. *Journal of the American Statistical Association*. 1986; 81:945–960.
- Jasjeet SS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization. *Journal of Statistical Software*. 2011; 42:1–52.
- Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *The Journal of the American Medical Association*. 2007; 298:1209–1212.

- Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*. 2011; 30:2601–2621. [PubMed: 21786278]
- Lo S, Singh K. The Product-Limit Estimator and the Bootstrap: Some Asymptotic Representations. *Probability Theory and Related Fields*. 1985; 71:455–465.
- Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, Daubert JP, Higgins SL, Brown MW, Andrews ML. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New England Journal of Medicine*. 2002; 346:877–883. [PubMed: 11907286]
- Rubin DB. Inference missing data. *Biometrika*. 1976; 63:581–592.
- Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*. 2007; 26:20–36. [PubMed: 17072897]
- Tung R, Zimetbaum P, Josephson ME. A critical appraisal of implantable cardioverter-defibrillator therapy for the prevention of sudden cardiac death. *Journal of the American College of Cardiology*. 2008; 52:1111–1121. [PubMed: 18804736]
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 2000.
- Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine-reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*. 2007; 357:2189–2194. [PubMed: 18032770]

Table 1

Illustration of TBR and THR using MADIT-II trial results. Numbers shown are two-year mortality (or survival) rates estimated by Kaplan-Meier method.

		ICD(Y_1)		
		Dead(0)	Alive(1)	Total
conventional	Dead(0)		TBR	21%
therapy (Y_0)	Alive(1)	THR		79%
Total		14%	86%	100%

Table 2

Covariates at baseline used in the analysis of the Multicenter Automatic Defibrillator Implantation Trial II (MADIT-II) trial.

Variable	Note	Value
Age	Age greater than 70 years	0:no, 1:yes
NYHA	New York Heart Association class > II	0:no, 1:yes
QRS	QRS interval greater than 120 milliseconds	0:no, 1:yes
AF	Atrial fibrillation at enrollment	0:no, 1:yes
BUN	50mg/dl > Blood urea nitrogen > 26 mg/dl	0:no, 1:yes

Table 3

Analysis results of the MADIT-II data. is the postulated odds ratio of the two potential binary outcomes conditional on X.

Estimate, 95%CI		Conventional medical therapy (N=425, # death=85)	ICD (N=643, # death=78)
2-year survival rate		0.79, (0.74–0.83)	0.86 (0.83–0.90)
TBR	= 1 (conditional independence)	0.18, (0.14 – 0.22) ^a ,	(0.14 – 0.22) ^b
	= 2	0.16, (0.12–0.20),	(0.12–0.20)
	= 3	0.15, (0.11–0.19),	(0.11–0.19)
	= 4	0.15, (0.11–0.19),	(0.11–0.19)
THR	= 1	0.10, (0.07–0.12),	(0.07–0.12)
	= 2	0.08, (0.06–0.11),	(0.06–0.11)
	= 3	0.07, (0.05–0.10),	(0.05–0.10)
	= 4	0.07, (0.05–0.09),	(0.04–0.09)

^aBased on normal approximation using asymptotic standard error estimate;

^bBased on normal approximation using standard error estimate based on 500 bootstrap samples.

Table 4

Results of 1000 Monte Carlo simulations each composed of 1000 subjects with a 3:2 allocation ratio for intervention and control arms.

	Odds ratio	True value	Bias	SE _M	SE _A	Cov. Prob
TBR	1	0.180	0.000	0.020	0.021	0.953
	2	0.166	0.001	0.020	0.020	0.961
	3	0.147	0.003	0.018	0.019	0.955
	4	0.133	0.003	0.018	0.018	0.956
THR	1	0.095	0.001	0.012	0.012	0.946
	2	0.087	0.001	0.012	0.012	0.942
	3	0.076	0.001	0.011	0.011	0.945
	4	0.068	0.003	0.011	0.011	0.948

Odds ratio: true marginal odds ratio of the two potential binary outcomes of survival beyond year 2 under intervention and control treatment; True value: true TBR/THR values; SEM: standard errors of estimators based on Monte Carlo data sets; SE_A: average of large sample standard error estimates; Cov. Prb: coverage probability of the 95% confidence interval using normal approximation with large-sample standard error estimate.