University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 264

The Impact Of Coarsening The Explanatory Variable Of Interest In Making Causal Inferences: Implicit Assumptions Behind Dichotomizing Variables

Ori M. Stitelman* Alan E. Hubbard† Nicholas P. Jewell‡

*Division of Biostatistics, UC Berkeley, ostitelman@berkeley.edu

†Division of Biostatistics, UC Berkeley, hubbard@berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper264

Copyright ©2010 by the authors.

The Impact Of Coarsening The Explanatory Variable Of Interest In Making Causal Inferences: Implicit Assumptions Behind Dichotomizing Variables

Ori M. Stitelman, Alan E. Hubbard, and Nicholas P. Jewell

Abstract

It is common in analyses designed to estimate the causal effect of a continuous exposure/treatment to dichotomize the variable of interest. By dichotomizing the variable and assessing the causal effect of the newly fabricated variable practitioners are implicitly making assumptions. However, in most analyses these assumptions are ignored. In this article we formally address what assumptions are made in dichotomizing variables to assess causal effects. We introduce two assumptions, either of which must be met, in order for the estimates of the causal effects to be unbiased estimates of the parameters of interest. We title those assumptions the Mechanism Equivalence and Effect Equivalence assumptions. Furthermore, we quantify the bias induced when these assumptions are violated. Lastly, we present an analysis of a Malaria study that exemplifies the danger of naively dichotomizing a continuous variable to assess a causal effect.

1 Introduction

A commonly used parameter of interest in observational studies designed to examine the effect of exposure, a, on outcome of interest Y, is $E[Y_a|V]$. This is interpreted as the mean outcome, Y, when every subject is set to a precise exposure, a, within a set level of observed baseline covariates, V. Variants of this parameter of interest are also commonly used throughout the causal inference literature. These parameters of interest are based on the causal inference framework originally proposed by Rubin [7], and are designed to evaluate the effect when the treatment, or exposure, is set to a specific level, a. Here, we examine the implications of these parameter of interest when the level of exposure, a, to which the subjects are set is really a range of values. Thus, a really represents many sub-levels of exposure. For example a binary indicator which describes smoking status classifies individuals as either smokers or non-smokers; however, among the smokers the amount which each individuals smoke varies and likewise among non-smokers the amount of second hand smoke exposed to varies. Thus, if one wanted to assess the causal effect of setting individuals to non-smoker the level non-smoker really is composed of many sub-levels of exposure. Such a hypothetical treatment is natural in a setting where one wants to evaluate the effect of lowering an exposure below a certain cut-off, as in the case of evaluating the effect of restricting air pollutants to a prespecified safe level.

The parameters of interest are functions of the observed distribution of the sub-levels of exposure within the level of a. We will refer to the distribution of the exposure as the treatment mechanism. It is important to consider two particular treatment mechanisms here. The first is the intended treatment mechanism, this mechanism is the distribution which sets the exposure to the levels desired by the intervention. The second is the implied treatment mechanism, this mechanism is the distribution of the exposure in the counterfactual world implied by the method used to estimate the parameter of interest. In the case of a true binary exposure there is only one possible treatment mechanism, the distribution in which all values are set to exposure level, a. However, in the case where one is interested in setting the exposure to a pre-specified cutoff or below, as in the case of the pollutant example above, there are an infinite number of possible distributions of the exposure. Thus, any distribution of the exposure whose maximum value is below the pre-specified cutoff is a possible treatment mechanism.

An example of a class of parameters of interest which incorporate $E[Y_a|V]$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

are proposed by Hubbard and van der Laan [4]. These parameters of interest are designed to quantify the effect of a treatment, or a set level of exposure, on a target population of interest. They are an extension of commonly used causal inference techniques and quantify the effect of an intended treatment on a population of interest. In addition to formulating a general approach for estimating a treatment effect they focus on casting these parameters as both a difference and ratio between the mean outcome under intervention and the mean outcome under the natural distribution of the treatment/exposure. These models are referred to as population intervention models. Specifically, they propose the two following classes of parameters for additive risk and relative risk:

$$\psi_{0,AR}(a,V) = E[Y_a|V] - E[Y|V] = m(a,V|\beta_0) \tag{1}$$

$$\psi_{0,RR}(a,V) = \frac{E[Y_a|V]}{E[Y|V]} = m(a,V|\beta_0) \tag{2}$$

for some Euclidean parameterization $\beta \rightarrow m(a,V|\beta)$.

In this paper we will use the population intervention model and its parameters of interest as an example as we explore the implications of dichotomizing the treatment variable as a way to estimate parameters which incorporate $E[Y_a|V]$. In section 2 we will examine the effect of dichotomizing the treatment variable and the traditional assumptions used in the causal inference framework. We will show that by dichotomizing the treatment/exposure the treatment mechanism implied by the method is the treatment mechanism observed in the data within level a. In section 3 two additional assumptions will be presented, which if either are not violated, an estimator for the parameter of interest will remain consistent under the intended treatment mechanism. These assumptions are (1) **Mechanism Equivalence** - the implied treatment mechanism is equal to the intended treatment mechanism and (2) **Effect Equivalence** - The expected outcome is the same across all sub levels of a. In section 4 the asymptotic bias which results from violations in the assumptions of section 3 will be examined. Finally, in section 5, an example is presented which examines the effect of parasite density on the recurrence of malaria.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

2 Dichitomizing The Treatment Variable

One possible estimation approach is to dichotamize the exposure at the specified level and then implement, designed for a binary exposure. Such an approach is a common way of implementing methods designed for binary exposures when dealing with exposures which are either categorical with more than two possible values or continuous. Some examples of this in the causal inference setting include Brotman et al. [1], Bryan et al. [2], Joffe et al. [5], and Tager et al. [8]. We will now examine the implied intervention mechanism when taking such an approach. More specifically, we will examine what intervening by setting A $= a$, when A is artificially dichotomized, suggests about the distribution of the sublevels of a after intervened upon population.

For this paper we will assume that exposure is experienced at discrete levels. For instace, let us examine what happens when exposure is initially randomized to five levels; let $A' \in \{1, 2, 3, 4, 5\}$ be the observed level of exposure for each subject and let $Y^{\Delta}(A')$ be the potential outcomes at each level of A'. Thus, the full data as defined in the counterfactual framework is $\left\{Y^{\Delta}(1), Y^{\Delta}(2), Y^{\Delta}(3), Y^{\Delta}(4), Y^{\Delta}(5), A'\right\}$ and the corresponding observed data is $\left\{Y^{\Delta}(A'), A'\right\}$. Define A as the dichotomized random variable which splits A' in the following way:

$$A = \left\{ \begin{array}{l} 1 \text{ if } A' \in \{1, 2\} \\ 0 \text{ if } A' \in \{3, 4, 5\} \end{array} \right\} \tag{3}$$

Finally, let $Y^*(A)$ be the potential outcomes at different levels of A. Thus, by implementing methods using only the dichotomized variable and the outcome it is as if we only observed $\{Y^*(A), A\}$ and considered the full data $\{Y^*(0), Y^*(1), A\}$ for each subject.

Two assumptions which are typically made in similar missing data problems will be useful here. The first is the consistency assumption:

$$Y = AY^*(1) + (1 - A)\, Y^*(0) \tag{4}$$

This assumption states that the observed outcome Y is equal to $Y^*(1)$ when treatment is set to A $= 1$ and to $Y^*(0)$ when the treatment is set

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

to A = 0. This assumption is commonly refered to as the Stable Unit Treatment Value Assumption (SUTVA)[7]. The second assumption is the randomization assumption:

$$A \perp \{Y^*(1), Y^*(0)\} \tag{5}$$

The randomization assumption states that the treatment level is set independent of the potential outcomes of the subject under the different levels of treatment.

We examine what happens to the mean outcome under the treatment, $E[Y_a]$, when we intervene by setting each individual to the exposure level $A = 1$ in the scenario when we implement a method on the data as if we only observed the dichotomized variable A:

$$\begin{aligned} E[Y_a] &= E[Y|A=1] \\ &\overset{CA}{=} E[AY^*(1) + (1-A)Y^*(0)|A=1] \\ &= E[Y^*(1)|A=1] \\ &\overset{RA}{=} E[Y^*(1)] \end{aligned} \tag{6}$$

where the second equality is the result of the consistency assumption and the last the result of the randomization assumption. Thus, the expected value of the observed outcome when the treatment level A is one is equal to the expected value of the potential outcomes for level A equal to 1.

By making two similar assumptions with regard to the treatment levels A' the $E[Y_a]$, the mean outcome when all subjects are set to level $A = 1$, can be evaluated under the true full data, corresponding to the scenario where A' is observed. The additional assumptions are another consistency assumption and randomization assumption for the exposure levels A'. The consistency assumption is:

$$Y = I(A'=1)Y^{\Delta}(1) + I(A'=2)Y^{\Delta}(2) + \ldots + I(A'=5)Y^{\Delta}(5) \tag{7}$$

and randomization assumption:

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

$$A' \perp \left\{ Y^{\Delta}(1), Y^{\Delta}(2), Y^{\Delta}(3), Y^{\Delta}(4), Y^{\Delta}(5) \right\} \tag{8}$$

These assumptions are as plausable as the corresponding assumptions under two treatment levels and may be stated in the same way as above but for five treatment levels instead of two. Returning to $E[Y_a]$ under these assumptions:

$$\begin{aligned}
E[Y_a] &= E[Y|A=1] \\
&\overset{CA}{=} E\Big[I(A'=1)\,Y^{\Delta}(1) + I(A'=2)\,Y^{\Delta}(2) + \\
&\quad \ldots + I(A'=5)\,Y^{\Delta}(5)\,|A=1\Big] \\
&= E\Big[I(A'=1)\,Y^{\Delta}(1) + I(A'=2)\,Y^{\Delta}(2)\,|A=1\Big] \qquad (9) \\
&\overset{RA}{=} p(A'=1|A=1)\,E\left[Y^{\Delta}(1)\right] + p(A'=2|A=1)\,E\left[Y^{\Delta}(2)\right] \\
&= \frac{p(A'=1)}{p(A=1)} E\left[Y^{\Delta}(1)\right] + \frac{p(A'=2)}{p(A=1)} E\left[Y^{\Delta}(2)\right]
\end{aligned}$$

Finally, combining the results from using the assumptions with respect to treatment levels A with the results for the assumptions with respect to treatment levels A':

$$\begin{aligned}
E[Y^*(1)] &= E[Y|A=1] \qquad (10) \\
&= \frac{p(A'=1)}{p(A=1)} E\left[Y^{\Delta}(1)\right] + \frac{p(A'=2)}{p(A=1)} E\left[Y^{\Delta}(2)\right]
\end{aligned}$$

Thus, by dichotomizing the exposure variable one implicitly assumes that the hypothetical intervention, setting individuals to $A = 1$, is achieved by randomly assigning individuals to treatment $A' = 1$ with probability $p(A' = 1)/p(A = 1)$ and to $A' = 2$ with probability $p(A' = 2)/p(A = 1)$. Thus the treatment mechanism employed assigns treatment to the population at a rate defined by the conditional distribution of A' given A.

Now let's extend these results to a situation where individuals are not randomly assigned treatment levels, as in the case of an observational study. Again we will consider the case where individuals are observed at treatment

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

levels A' and all analysis is conducted using the dichotomized treatment level A as done above. However, additional covariates, W, are also observed. The randomization assumption is no longer valid because individuals are not randomly assigned a level of treatment and thus there may be an association between the potential outcomes and the level of exposure one experiences. In such instances the assumption of no unmeasured confounders may be plausable.

$$A \perp \{Y^*(1), Y^*(0)\} | W \tag{11}$$

In words this assumption means that treatment level is assigned independent of the potential outcomes given the confounders, V. Using these assumptions, the following links the observed data using the exposure levels A to the counterfactual outcomes under levels of A:

$$\begin{aligned} E_W[E[Y_a|W]] &= E_W[E[Y|A=1,W]] \\ &\overset{CA}{=} E_W[E[AY^*(1)+(1-A)Y^*(0)|A=1,W]] \\ &= E_W[E[Y^*(1)|A=1,W]] \\ &\overset{RA}{=} E_W[E[Y^*(1)|W]] \\ &= E[Y^*(1)] \end{aligned} \tag{12}$$

One additional assumption is necessary for the above to be true because it is important that we are not conditioning on an event that will occur with probability zero. The necessary assumption is The Experimental Treatment Assumption (ETA)[6]. ETA insures that every possible value of A has a positive probability of occuring regardless of the level of baseline covariates, W. Equivalently, $p(A=1|W)$ and $p(A=0|W)$ are bounded away from zero almost everywhere. Extending the no unmeasured confounders assumption to the levels of A' as follows,

$$A \perp \left\{Y^\Delta(1), Y^\Delta(2), Y^\Delta(3), Y^\Delta(4), Y^\Delta(5)\right\} | W, \tag{13}$$

similarly yields the following in terms of levels of the exposure A'

$$E_W[E[Y_a|W]] = E_W[E[Y|A=1,W]]$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

$$
\begin{aligned}
&= E_W\left[E\left[I\left(A'=1\right)Y^{\Delta}\left(1\right)+I\left(A'=2\right)Y^{\Delta}\left(2\right)+\right.\right.\\
&\quad \left.\left.\ldots+I\left(A'=5\right)Y^{\Delta}\left(5\right)|A=1,W\right]\right] \qquad (14)\\
&= E_W\left[E\left[I\left(A'=1\right)Y^{\Delta}\left(1\right)+I\left(A'=2\right)Y^{\Delta}\left(2\right)|A=1,W\right]\right]\\
&= E_W\left[p\left(A'=1|A=1,W\right)E\left[Y^{\Delta}\left(1\right)|W\right]+p\left(A'=2|A=1,W\right)E\left[Y^{\Delta}\left(2\right)|W\right]\right]
\end{aligned}
$$

Combining the two results:

$$
\begin{aligned}
E\left[Y^{*}\left(1\right)\right] &= E_W\left[E\left[Y_a|W\right]\right]\\
&= E_W\left[p\left(A'=1|A=1,W\right)E\left[Y^{\Delta}\left(1\right)|W\right]+\right. \qquad (15)\\
&\quad \left.p\left(A'=2|A=1,W\right)E\left[Y^{\Delta}\left(2\right)|W\right]\right]
\end{aligned}
$$

Again, the treatment mechanism employed assigns treatment to the population at a rate defined by the conditional distribution of A' given A; however, now it is within levels of W.

Typically, one is interested in dichotomizing a continuous random variable, let us call this variable A''. We will take a hueristic approach to address this issue. In such a scenario mapping the continuous variable, A'', into a categorical variable, A' with an arbitrary amount of levels may help in considering this set up. For fine enough cut points, the discrete conditional distribution of $A'|A,W$ is an approximation of $A''|A,W$ and in most situations, it is reasonable to assume that $E\left[Y|A'',W\right]$ is equivalent across A'' within each level of A'. Under these two conditions the the results for the categorical mapping into a dichotomous variable above may be extended to dichotomizing the continuous case.

3 Intended vs. Implied Treatment Mechanism

This section evaluates the difference between the treatment mechanism induced by dichotomizing the exposure variable and a possible intended treatment mechanism. In order to examine the difference of these treatment mechanisms reconsider the five level treatment assignment of A' in the previous section where one observes covariates, V, and individuals are not randomly assigned to an exposure level. However, now instead of there being

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

five levels of A' let there be an arbitrary number of levels, k, of A', where the first j levels of A' are mapped into the level $A = 1$. Let $p^*(A'|V)$ equal the intended probability of being assigned to level A' given V under the proposed intervention, and $E[Y_{a^*}|V]$ is the expected value of the outcome given the intended treatment mechanism within strata, V. The values at which $p^*(A'|V)$ are set should be thought of as attainable probabilities of exposure at each level given the intended course of intervention.

The population intervention model as presented by Hubbard and van der Laan presents an estimator which is consistent for the parameters of interest presented in equations 1 and 2. These parameters of interest include $E[Y_a|V]$ as a component. Furthermore, the estimators are consistent under the implied treatment mechanism. As presented in the previous section, the implied treatment mechanism, by dichotomizing the random variable, is to assign treatment to the population at a rate defined by the conditional distribution of A' given A within strata of V. Thus, the estimators are a consistent estimate of the following parameters under additive and relative risk:

$$\begin{aligned} E[Y_a|V] - E[Y|V] &= p(A' = 1|A = 1, V)\, E\left[Y^{\Delta}(1)|V\right] + \ldots \\ &\quad + p(A' = k|A = 1, V)\, E\left[Y^{\Delta}(k)|V\right] - E[Y|V] \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{E[Y_a|V]}{E[Y|V]} &= \frac{p(A' = 1|A = 1, V)\, E\left[Y^{\Delta}(1)|V\right]}{E[Y|V]} + \ldots \\ &\quad + \frac{p(A' = k|A = 1, V)\, E\left[Y^{\Delta}(2)|V\right]}{E[Y|V]} \end{aligned} \tag{17}$$

These parameters of interest under the implied treatment mechanism coincide with the parameter of interest where the intervention is the intended treatment mechanism if either of the following two scenarios is true:

1. MECHANISM EQUIVALENCE: The intended mechanism after intervention coincides with the observed mechanism which results by setting A equal to 1. The conditional probability of A' given A within

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

strata of V is equivalent to the intended intervention probabilities, $p^*(A'|V)$, for all A' mapped into the intervention level A.

$$p^*(A'|V) = p(A'|A=1,V)\,\forall A' \in \{1,\ldots,j\} \tag{18}$$

2. EFFECT EQUIVALENCE: The effect of A is equivalent within strata of V for the all levels of A'. The expected value of the potential outcomes given V and A' is equivalent for all A' within a particular strata, V.

$$E\left[Y^{\Delta}(1)|V\right] = \ldots = E\left[Y^{\Delta}(j)|V\right] \tag{19}$$

Thus, by dichotomizing the exposure variable and implementing the population intervention model, assuming either that the intervention will set the probability of exposure level to the conditional probabilities observed in the data or the expected value of the outcome is the same across the different levels within A for a given V, the estimator will be consistent for the parameter of interest under the intended treatment mechanism. If either of these two assumptions are plausable, the population intervention model implemented using a dichotomized treatment variable is a reasonable estimator that will produce consistent estimates of well-defined parameter of interest. In the following section, we will examine ways to implement the population intervention model when either of these two assumptions are not reasonable.

4 When Assumptions Are Not Plausable

In this section we will explore what to do in two situations when there is a possibility that the assumptions presented in the previous section do not hold. The first situation is one in which the data is dichotomized because the underlying variable A' is not observed either at finer cut-offs or as a continuous variable below a certain cut-off. The second situation is one in which the data is observed at finer levels within both $A=0$ and $A=1$.

In some instances, as in the case of most variables which measure exposure to pollutants, the level of exposure, A', may only be determined above a certain level. All other subjects are designated as having an exposure below the detectable limit. A natural intervention in such cases is to examine the counterfactual world where all individuals have their exposure reduced to a level below the detectable level, corresponding to $A = 1$(below the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

detectable level) and $A = 0$ (above the detectable level). Thus, the theoretically different levels of A' which map into $A = 1$ are not observed. Short of estimating the levels of A' given the covariates, V, which is not a trivial problem and most likely involves non-testable parametric assumptions since one is estimating values outside their observed range, making one of the two assumptions from the following section and implementing the population intervention model on the dichotomized variable may be the most reasonable course of action. However, it is important in those situations to be aware of these assumptions and gauge their plausability on a case by case basis.

In other situations where either of the two assumptions of the previous section are not plausable and A' is observed at a finer level within $A = 1$, extensions of the methods presented by Hubbard and van der Laan that do not rely on a binary exposure variable are possible. We will now present a method of using the population intervention model to estimate a parameter of interest under the intended treatment mechanism when either of the two necessary assumptions are not valid.

Let's reconsider the scenario in section two where there are five observed levels of A', covariates V are observed, and exposure level is as it would be in an observational study, i.e., not randomized. Additionaly, let p be a vector of the probabilities that make up the intended mechanism, p_1 and p_2, which are equal to $p^* (A' = 1|A = 1, V)$ and $p^* (A' = 2|A = 1, V)$, respectively. Given that the parameter of interest is additive risk, one would be interested in a consistent estimate of the following:

$$E\left[Y_{a^*}|V\right] - E\left[Y|V\right] = p_1 E\left[Y^{\Delta}(1)|V\right] + p_2 E\left[Y^{\Delta}(2)|V\right] - E\left[Y|V\right] \quad (20)$$

Hubbard and van der Laan present consistent estimates , $m(a, V|\widehat{\beta})$ of $m(a, V|\beta) = E\left[Y_a|V\right] - E\left[Y|V\right]$ for a given level of treatment, a. As a result, the following is a consistent estimate of $E\left[Y_{a^*}|V\right]$:

$$p_1 \left\{E\left[Y|V\right] + m\left(1, V|\widehat{\beta}\right)\right\} + p_2 \left\{E\left[Y|V\right] + m\left(2, V|\widehat{\beta}\right)\right\} \quad (21)$$

Where $m\left(1, V|\widehat{\beta}\right)$ and $m\left(2, V|\widehat{\beta}\right)$ are estimated as suggested in Hubbard and van der Laan for treatment levels $A' = 1$ and $A' = 2$ respectively. The estimate for $E\left[Y_{a^*}|V\right]$ can be rewritten in the following way:

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

$$(p_1 + p_2)\{E[Y|V]\} + p_1 m(1, V|\beta) + p_2 m(2, V|\beta) \tag{22}$$

Finally, the parameter of interest, $E[Y_{a^*}|V] - E[Y|V]$, may be estimated by the following consistent estimate, $p_1 m\left(1, V|\widehat{\beta}\right) + p_2 m\left(2, V|\widehat{\beta}\right)$.

These results may be generalized in a similar way to the relative risk parameter of interest and the general population intervention model parameter of interest presented by Hubbard and van der Laan. So by consistently estimating $m(A' = 1, V|\beta)$ and $m(A' = 2, V|\beta)$ and combining those estimates using p, consistent estimates of the parameter of interest under the intended treatment mechanism may be achieved.

Additionally, the asymptotic bias of using the above estimates can be expressed in the following way in terms of the potential outcomes at different levels of A', the intended treatment mechanism, p, and the conditional probabilities of the $A'|A, V$, which make up the implied treatment mechanism due to dichotomizing the exposure variable:

$$\begin{aligned} &\{p_1 - p(A' = 1|A = 1, V)\} E\left[Y^{\Delta}(1)|V\right] + \ldots \\ &\quad + \{p_k - p(A' = k|A = 1, V)\} E\left[Y^{\Delta}(k)|V\right] \end{aligned} \tag{23}$$

Alternatively, the asymptotic bias can be written in terms of only the intended treatment mechanism, p, the potential outcomes at levels of A', and the potential outcomes at levels of A:

$$p_1\left\{E\left[Y^{\Delta}(1)|V\right] - E\left[Y^*(1)|V\right]\right\} + \ldots + p_k\left\{E\left[Y^{\Delta}(k)|V\right] - E\left[Y^*(1)|V\right]\right\} \tag{24}$$

The first formulation of the bias, equation 22, illustrates the effect of the difference between the intended treatment mechanism and the hypothetical treatment on the bias. Violations in the mechanism equivalence assumption will result in larger values of bias since these violations will result in the $\{p_i - p(A' = i|A = 1, V)\}$ terms being further away from zero. Again, it is clear that if the intended treatment mechanism equals the implied treatment mechanism from dichotomizing the estimate is consistent. The second formulation, equation 23, illustrates the effect of violations from the second assumption, effect equivalence, on the bias. Thus, when there are large differences in the expected potential outcomes at levels A' and the expected potential outcomes A there will be large bias.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Figures 1 and 2 below illustrate the effects of violations of the proposed assumptions on asymptotic bias. For the purpose of these figures let us assume that the level of A equal to 1, to which all subjects would be set in the hypothetical treatment, is truly divided into two distinct levels, $A' = 1$ and $A' = 2$. Since there are only two levels of A' the intended intervention probabilities at each level of V are fully defined by one probability, p_1. Thus, the intended treatment mechanism involves setting within strata of V, A' equal to 1 with probability p_1 and to $A' = 2$ with probability $1 - p_1$. For Figure 1, $E\left[Y^{\Delta}(1)|V\right] = 2$, and $E\left[Y^{\Delta}(2)|V\right] = 3$, and the true conditional probability $p(A' = 1|A = 1, V) = .45$. Figure 1 shows that when the intended intervention probabilities, p_1, are equivalent to the true conditional probabilities, $p(A' = k|A = 1, V)$, the asymptotic bias is equal to zero. Furthermore, as the intended intervention probabilities deviate from the true conditional probabilities the asymptotic bias increases in magnitude in a linear fashion. In fact, the slope of the line is equal to $E\left[Y^{\Delta}(1)|V\right] - E\left[Y^{\Delta}(2)|V\right]$, so for a $p_1 - p(A' = k|A = 1, V)$ deviation from the observed conditional probability there will be a $\{p_1 - p(A' = k|A = 1, V)\}\left\{E\left[Y^{\Delta}(1)|V\right] - E\left[Y^{\Delta}(2)|V\right]\right\}$ change in the asymptotic bias.

Figure 2 similarly illustrates the effect of violations in the Effect Equivalence assumption on asymptotic bias. For this figure the values were set to $E\left[Y^{\Delta}(1)|V\right] = 2$, $p_1 = .2$, and $p(A' = 1|A = 1, V) = .8$. When $E\left[Y^{\Delta}(1)|V\right]$ is equal to $E\left[Y^{\Delta}(2)|V\right]$ the asymptotic bias is equal to zero. Once again the asymptotic bias has a linear relationship with respect to deviations from the proposed assumption. For a unit deviation from effect equivalence there is a $p_1 - p(A' = k|A = 1, V)$ deviation in asymptotic bias.

In both figures it is clear that the overall size of the asymptotic bias is dependent on the extent to which both assumptions are violated. Recall that the asymptotic bias is equal to zero if either the Effect Equivalence or Mechanism Equivalence assumption is satisfied. The slope of Figure 1, which depicts the effect of deviations in the treatment mechanism, is equal to the deviation in the expected value of the outcome within levels of A'. Likewise, the slope of Figure 2, which depicts the effect of deviations in the expected value of the outcome, is equal to the deviation in the treatment mechanism. Thus, the larger the difference between the intended treatment mechanism and the true conditional probability the larger the effect of deviations in expected outcome, and vice versa.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

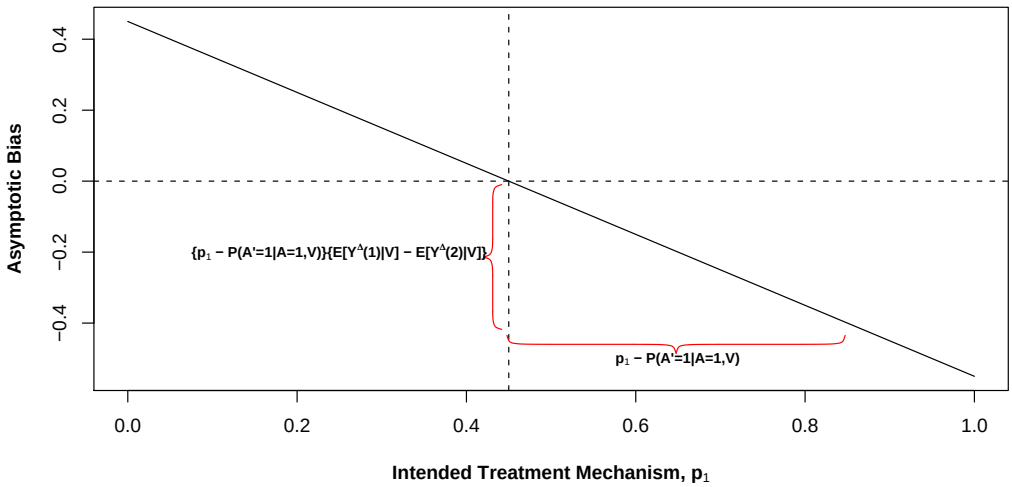

Figure 1: Asymptotic Bias vs Intended Treatment Mechanism

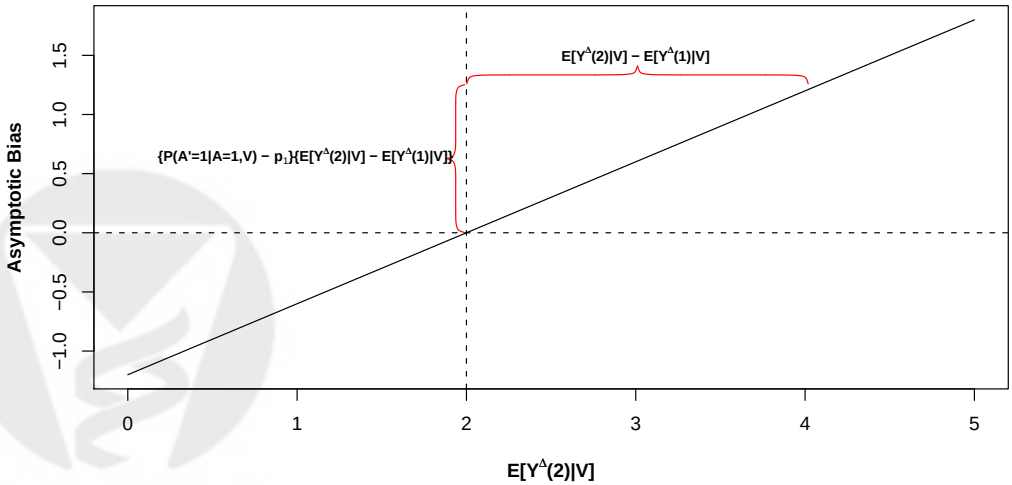

Figure 2: Asymptotic Bias vs Expected Value of Potential Outcome

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

5 Example: Effect Of Parasite Density On Malaria Recurrence

We will now examine how dichotomizing the exposure effects estimates of the population intervention model additive risk when measuring the effect of baseline parasite density on the recurrence of Malaria.[1] The following analysis was motivated by data used in Greenhouse et al. 2006[3]. The exposure, baseline parasite density, is measured on a continuous scale between 2,000 and 1,000,000 parasites/μL and the observed outcome Y is a binary variable indicating whether or not an individual had a Malaria recurrence. A hypothetical intervention which will reduce baseline parasite density below 39,000 will be examined.

The parameter of interest is the additive risk parameter, $E[Y_a] - E[Y]$ for a binary variable A equal to 1 when parasite density is less than 39,000 and 0 otherwise. As shown above this parameter implies a treatment mechanism which sets individuals to parasite density levels below 39,000 at a rate equal to the conditional probability of the parasite density given that the parasite density is below 39,000. However, the designer of the intervention claims that it is able to reduce baseline viral load to less than 5,000 for 70 percent of patients and between 5,001 and 39,000 for the other 30 percent independent of the patients other characteristics. Thus, the intended parameter of interest is $E[Y_{a^*}] - E[Y]$, where a^* is the intervention which sets individuals to less than 5,000 UNITS 70 percent of the time and between 5,001 and 39,000 UNITS 30 percent of the time. Table 1 presents the estimates of the parameter $E[Y_a] - E[Y]$ for different levels of a using G-computation.

It is clear from this table that the Effect Equivalence assumption is violated since the effect is not the same among the less than 5,000 group and the 5,000 to 39,000 group. Also, about 15 percent of the individu-

[1] For this example G-computation methods will be used to estimate the parameter of interest, in Hubbard and van der Laan double robust estimates of the parameters of interest are also presented. In certain situations the double robust estimator exhibits nicer properties, namely they are consistent if either the model of the outcome is consistent or the model of the probability of the treatment given the covariates is consistent. However, for the sake of this example the simplicity of the G-computation estimator allows us to more clearly demonstrate our focus on the effects of dichotomizing the exposure/treatment variable without getting into the details of double robust estimation. Furthermore, for simplicity, we will examine the effect over the entire population and not for a specific subpopulation, V.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

Table 1: Estimate of Additive Risk For Different Exposure Levels, a

Exposure, a	$E[Y_a] - E[Y]$
$< 5,000$	-.073
$5,000 - 39,000$	-.032
$< 39,000$	-.039

als with parasite density less than 39,000 have a parasite density less than 5,000 far less than the 70 percent of individuals which hypothetical treatment intends to lower to less than 5,000 indicating that the Mechanism equivalence assumption is also violated. The estimate of the parameter of interest, $E[Y_{a^*}] - E[Y]$, for the hypothetical intervention is -.061, or 56 percent larger than the estimate obtained from simply dichotomizing the treatment variable at 39,000 and estimating the parameter. This example nicely illustrates how different ones estimate of the parameter of interest may be from what one intended to estimate if they naively dichotomize the exposure/treatment variable without further exploration of the effect within sub-levels of the intended treatment and without consideration for the distribution of the exposure which may be achieved by the proposed intervention.

6 Discussion

The above results illustrate that by estimating the parameters of the population intervention model for the different levels of A' within $A = 1$ separately and then combining those estimates using the user-supplied intended treatment mechanism, p, one can arrive at a consistent estimate of the parameter of interest. Furthermore, this estimate will exhibit the same properties as the estimates of Hubbard and van der Laan, 2008. Specifically, these estimates will be remain unbiased when the nuisance parameters are misspecified and efficient when they are specified correctly.

The assumptions which are necessary to have consistency in section two, mechanism equivalence and the effect equivalence, are not specific to the population intervention model. In fact, at least one of the two assumptions is necessary in any method which provides consistent estimates of a parameter of interest which includes $E[Y_a|V]$. Assessing the plausability of one

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

or both of these assumptions is a necessary consideration when deciding to dichotomize an exposure variable and implement accepted methods for a naturally binary exposure. Even in situations when the exposure variable initially appears to be naturally binary, considering the plausability of these assumptions may be reasonable. For example, when the exposure A is defined as equal to one for smokers and equal to zero for nonsmokers and one is interested in the counterfactual world where the entire population is set to non-smoker, evaluating the plausability of these assumptions is a reasonable course of action. Cigarette smoke exposure is trully a continuous random variable and the level of exposure is not equal among all smokers, and likewise is not equal among all non-smokers; thus, by using the dichotomous exposure variable, most likely the result of self report, one must either assume that an intervention will set all subjects to the smoke exposure levels in the nonsmoker group at a rate equal to the conditional probability of being nonsmoker given true cigarette exposure and covariates, V (Mechanism Equivalence) or that the effect at all cigarette smoke levels within the nonsmoker group is the same (Effect Equivalence). Furthermore, these assumptions are not specific to a binary random variable and should be considered in all situations when another random variable is mapped into a categorical variable or in which the data is collected at a coarser level than the true exposure.

References

[1] Rebecca M Brotman, Mark A. Klebanoff, Tonja R. Nansel, and William W. Andrews. A longitudinal study of vaginal douching and bacterial vaginosis—a marginal structural modeling analysis. *American Journal of Epidemiology*, 168(2):188–196, May 2008.

[2] Jenny Bryan, Zhuo Yu, and Mark J. van der Laan. Analysis of longitudinal marginal structural models. *Biostatistics*, 5(3):361–380, 2004.

[3] B. Greenhouse and A Myrick et al. Validation of microsatellite markers for use in genotyping polyclonal plasomodium falciparum infections. *Am J Trop Med Hyg.*, 75(5):836–842, 2006.

[4] Alan E Hubbard and Mark J. van der Laan. Population intervention models in causal inference. *Biometrika*, 95(1):35–47, July 2008.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper264

[5] Marshall M. Joffe and Thomas R.Tenhave. Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician*, 58(4):272–279, November 2004.

[6] R Neugebauer and M. J. van der Laan. Why prefer double robust estimates? illustration with causal point treatment studies. *bepress*, 2002.

[7] D. B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annal of Statistics*, 6:34–58, 1978.

[8] Ira B Tager, Thaddeus Haight, Barbara Sternfeld, Zhuo Yu, and Mark J. van der Laan. Effects of physical activity and body composition on functional limitation in the elderly. *Epidemiology*, 15(4):479–493, July 2004.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press