

What Can Be Learned from a Simple Table? Bayesian Inference and Sensitivity Analysis for Causal Effects from 2×2 and $2 \times 2 \times K$ Tables in the Presence of Unmeasured Confounding *

Kevin M. Quinn[†]

This Version: February 17, 2009
Original Version: January 11, 2008

Abstract

What, if anything, should one infer about the causal effect of a binary treatment on a binary outcome from a 2×2 cross-tabulation of non-experimental data? Many researchers would answer “nothing” because of the likelihood of severe bias due to the lack of adjustment for key confounding variables. This paper shows that such a conclusion is unduly pessimistic. Because the complete data likelihood under arbitrary patterns of confounding factorizes in a particularly convenient way, it is possible to parameterize this general situation with four easily interpretable parameters. Subjective beliefs regarding these parameters are easily elicited and subjective statements of uncertainty become possible. This paper also develops a novel graphical display called the *confounding plot* that quickly and efficiently communicates *all* patterns of confounding that would leave a particular causal inference relatively unchanged.

*An earlier version of this paper was awarded the 2008 Gosnell Prize for Excellence in Political Methodology. I thank Adam Glynn, Dan Ho, Luke Keele, Gary King, and seminar participants at Penn State University and the Northeast Methods Meeting for helpful comments and conversations. I gratefully acknowledge support from the U.S. National Science Foundation (grant BCS 05-27513) and the Institute for Quantitative Social Science at Harvard University. Software to perform the analyses described in this paper is available at <http://cran.r-project.org/web/packages/SimpleTable/index.html>.

[†]Associate Professor, Department of Government and The Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. kevin_quinn@harvard.edu

1 Introduction

What, if anything, should one infer about the causal effect of a binary treatment X on a binary outcome Y from a 2×2 cross-tabulation of non-experimental data? As an example, consider Table 1 which presents data that might be relevant to test the so-called “jury aversion” hypothesis (Oliver and Wolfinger, 1999). This table presents data on citizens’ perceived source of jury lists ($X = 0$: sources of jury lists do not include voter lists, $X = 1$: sources of jury lists do include voter lists) and voter registration ($Y = 0$: citizen did not register to vote, $Y = 1$: citizen did register to vote) from a study by Oliver and Wolfinger (1999). One might be tempted to infer that since the estimated

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	19	143
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	114	473

Table 1: *Perceived Source of Jury Lists and Voter Registration Among Citizens With Some Self Reported Knowledge of How Jury Lists are Constructed.* Each entry is the number of citizens in that category. Data from Table 2 of Oliver and Wolfinger (1999). An additional 639 citizens (46% of total) responded that they did not know the source of jury lists.

probability of registering to vote given $X = 0$ equals $143/(19 + 143) \approx 0.88$ and the estimated probability of registering to vote given $X = 1$ equals $473/(114 + 473) \approx 0.81$, perceiving voter lists to be a source of jury lists decreases the probability of registering to vote (among those with some knowledge of how jury lists are formed) by about 7 percentage points. Is this a reasonable inference?

Most scholars would argue that this would not be a reasonable inference. Since it seems likely that many variables that are common causes of both X and Y —such as occupation, education, civic involvement, etc.—are omitted,¹ the naive estimate of a 7 percentage point effect is likely biased.

¹Such omitted variables are often referred to as *confounders*. When it is not possible to measure all such confounders some say that *unmeasured confounding* is present. Rigorous definitions of these concepts are somewhat

Many researchers would go on to say that not much of anything could be said about causality from this table because of this bias. While there are certainly aspects of truth to the position that unmeasured confounding invalidates many attempts at causal inference, it turns out that it is possible to learn some things about the size of the causal effects of interest regardless of the nature and degree of the unmeasured confounding. For instance, Manski (1990) derived bounds for the average treatment effect under very general assumptions and showed that, in the case of binary treatment and outcome, the width of these bounds is 1.² Since the average treatment effect can take values between -1 and 1, Manski's bounding interval always includes the null value of no effect. Manski (2003) has also shown how auxiliary assumptions can narrow this bounding interval.

While this is clearly important work in that it provides a very precise statement of what the observed data alone can tell one about the causal effects of interest, it does not go quite as far as many social scientists would like on two counts. First, it provides absolutely no guidance as to where within the bounding interval the causal effect of interest is likely to be. Some may believe that all values within the interval are equally likely but, as we will see later, such a belief will often be inconsistent with reasonable beliefs about the nature of the unmeasured confounding. Second, the standard approach pursued in Manski (1990) is a large sample approach that does not account for sampling variability.³ This makes the interpretation of bounds derived for sparse tables somewhat unclear.

This paper address these issues by doing the following. It reduces the case of arbitrary unmeasured confounding with binary treatment and binary outcome to its most basic, yet still general, form. It does this in a way that provides a readily interpretable parameterization of the key quantities. It then shows how a Bayesian prior distribution can be placed over the four free parameters that govern the type and extent of the unmeasured confounding. If one is willing and able to use background knowledge to make some (possibly weak) assumptions about the nature of the unmeasured confounding, interested readers should consult Pearl (2000).

²For related work see Robins (1989); Manski (1993, 2003); Imai and Yamamoto (2008) and Balke and Pearl (1997).

³But see Imai and Soneji (2007) for an approach based on the bootstrap that does account for sampling variability.

sured confounding, sharp posterior estimates of causal effects are easy to calculate. Since these assumptions are formalized within the Bayesian framework, subjective uncertainty about causal effects is calculated in a logically coherent manner. The end result is a procedure that allows researchers to make probability statements about the likely size of causal effects based on the evidence in a 2×2 (or $2 \times 2 \times K$) table regardless of the sample size and the amount of unmeasured confounding. It thus directly addresses the shortcomings of the traditional bounding analysis.

The reasons for developing this approach to analyzing tabular data are three-fold. First, 2×2 tables are widely used (particularly in older work) and it would be useful to know how a rational person should interpret the information in these tables. Second, once the simple 2×2 case is understood it is easy to extend the main ideas in this paper to more complicated settings with binary treatment and outcome and measured categorical confounders. Interestingly, while conditioning on a measured confounder may make prior elicitation easier it does not change the large-sample nonparametric bounds on the average treatment effect. Thus the benefits of adjusting for multiple measured confounders when substantial *unmeasured* confounding is believed to be present may be fewer than many researchers seem to believe. Finally, to the extent that many political scientists are primarily interested in the sign of hypothesized causal effects, a well-designed (Bayesian) sensitivity analysis may provide them with stronger evidence for or against their hypothesis than the results from a small number of regression or matching analyses. Indeed, this was the motivation for what is commonly believed to be the first sensitivity analysis (Cornfield et al., 1959).⁴

The approach presented in this paper is related to, but distinct from, many sensitivity analysis methods that can be used to see how particular departures from the experimental ideal impact point estimates of causal effects. Typically, such approaches attempt to directly model the confounder-outcome association and the confounder-treatment association (Schlesselman, 1978; Rosenbaum and Rubin, 1983a; Lin et al., 1998). Such approaches require the practitioner to make assumptions about (a) the nature of the confounding variable (whether it is univariate/multivariate, dis-

⁴See Lin et al. (1998), p. 948.

crete/continuous, etc.) as well as (b) the association between this assumed confounder and outcomes and treatment status. Needless to say, it is very difficult for most practitioners to think coherently about the nature of the unmeasured confounder(s) and the resulting (conditional) associations. Other approaches reduce the sensitivity analysis to an examination of changes in a scalar parameter that can be varied by the researcher Rosenbaum (1987, 2002a,b). While this makes it much easier to move forward with a sensitivity analysis, the interpretation of this scalar parameter is not always as transparent as it might seem (Robins, 2002). The approach developed in this paper is at least as (sometimes more) general than the approaches described above and it also only requires researchers to specify their beliefs over easily interpreted quantities.

The work that is most similar to this paper is Chickering and Pearl (1997). That work relies on similar ideas to develop a Bayesian approach to the related problem of analyzing experimental data with non-compliance. A major difference between the two approaches is that the data of interest for this paper admit a much simpler factorization that makes prior elicitation and inference much more straightforward. A much smaller set of free parameters need be considered here and Markov chain Monte Carlo methods are not necessary for inference. This makes it possible to write easy-to-use open source software for prior elicitation and inference.⁵ Thus, the methods in the current paper are much more likely to be successfully used by empirical social scientists. Relatedly, the structure of the current problem allows for a novel graphical display called the *confounding plot* that (a) is quick and easy to compute, and (b) shows the entire range of possible types of confounding that would not appreciably change a given causal inference. This simple graph is so easy to generate and so informative that there is no reason it should not be a part of every analysis that attempts to make causal inferences from a 2×2 table.

The paper proceeds as follows. In Section 2 we introduce the necessary terminology and notation and then show how posterior inferences can be constructed from a 2×2 table with general unmeasured confounding. Section 3 shows how causal quantities of interest such as the average

⁵This software takes the form of an R (R Development Core Team, 2007) package named `SimpleTable`. This package is available at <http://cran.r-project.org/web/packages/SimpleTable/index.html>.

treatment effect and the relative risk can be written in terms of the model parameters from Section 2. Large sample nonparametric bounds on these causal quantities are also derived in this section. These bounds coincide with those of Manski (1990) although the derivation is slightly different. Section 4 discusses the choice of prior distribution for the model parameters and then describes a simple posterior sampling algorithm that does not require Markov chain Monte Carlo. Section 5 describes the construction and interpretation of the novel confounding plot discussed above. In Section 6 we revisit the example data in Table 1. Here we see how defensible prior beliefs can be operationalized in a prior distribution over the model parameters and what this implies for inferences about a possible jury aversion effect. The final section concludes.

2 Terminology, Notation, and the Probability Model

Consider the situation in which one is interested in assessing the causal effect of a binary treatment variable on a binary outcome. We let $X_i \in \{0 \text{ (control)}, 1 \text{ (treatment)}\}$ denote the treatment status of unit i , $Y_i \in \{0 \text{ (failure)}, 1 \text{ (success)}\}$ denote the observed outcome from unit i , and $i = 1, \dots, n$ index individual units. Throughout the rest of the paper we will assume that (X_i, Y_i) for $i = 1, \dots, n$ are independent replicates drawn from some joint distribution P_{XY} . We use the notation x_i and y_i to denote individual realizations of X_i and Y_i and \mathbf{x} and \mathbf{y} to denote the n -vectors of these realized values.

We adopt a counterfactual causal model that is consistent with the work of Rubin (1974, 1978); Robins (1986) and Pearl (1995, 2000) (see also Holland (1986) for a general review and Morgan and Winship (2007) for an introductory treatment from a social science perspective). We adopt the notation $Y_i(X_i = 0)$ and $Y_i(X_i = 1)$ to denote the value of unit i 's outcome variable if X_i is set equal to 0 and 1 respectively by an outside intervention that leaves all pre-intervention variables unchanged. Unit-level causal effects of changing X from 0 to 1 on Y in unit i are defined in terms of comparisons of $Y_i(X_i = 0)$ and $Y_i(X_i = 1)$. Since unit i cannot simultaneously receive both treatment and control, one of $Y_i(X_i = 0)$ and $Y_i(X_i = 1)$ is a counterfactual quantity.

Throughout the rest of this paper we will refer to $Y_i(X_i = 0)$ and $Y_i(X_i = 1)$ as *potential outcomes* or *counterfactual outcomes*.

Rather than focusing on unit-level causal effects, we will concern ourselves with aggregate effects within some collection of units. Here, interest centers on the *post-intervention distribution* $\Pr(Y(X = x) = y)$ for $x = 0, 1$ and $y = 0, 1$. The post-intervention probability $\Pr(Y(X = x) = y)$ gives the probability that the outcome variable from a randomly chosen unit will take value y when the unit in question is assigned $X = x$.

Causal quantities of interest such as the average treatment effect

$$ATE = \Pr(Y(X = 1) = 1) - \Pr(Y(X = 0) = 1)$$

and the relative risk

$$RR = \frac{\Pr(Y(X = 1) = 1)}{\Pr(Y(X = 0) = 1)}$$

can be defined in terms of the post-intervention distribution.

Note that the post-intervention distribution depends on counterfactual outcomes and is generally not equivalent to $P(Y = y|X = x)$. This conditional distribution can be calculated directly from P_{XY} . This latter distribution is sometimes called the *pre-intervention distribution*. While the pre-intervention distribution P_{XY} can be consistently estimated without maintaining untestable assumptions, the post-intervention distribution can only be estimated consistently if untestable causal assumptions are maintained (Rubin, 1978; Holland, 1986; Robins, 1986; Pearl, 1995, 2000).

Specifically, if one is willing to assume that treatment assignment is strongly ignorable⁶ and what Rubin (1980, 1986) has called the stable unit treatment value assumption (SUTVA)⁷ holds, then the post-intervention probabilities are given by the pre-intervention conditional probabilities

⁶Formally, strong ignorability of treatment assignment means that $[Y(X = 0), Y(X = 1)] \perp\!\!\!\perp X$. In words, the (counterfactual) joint distribution of potential outcomes is independent of treatment. Somewhat more intuitively, strong ignorability of treatment assignment means that a unit's potential outcomes do not depend on whether the unit is assigned to treatment or control. Note that this *does not* imply that the observed outcome Y is independent of treatment X .

⁷SUTVA has two parts. To quote Rubin (1986): "SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive, and this holds for [all units and all treatments]" (p. 961).

of Y given X :

$$\Pr(Y(X = x) = y) = \Pr(Y = y|X = x). \quad (1)$$

Estimates of $\Pr(Y = y|X = x)$ and hence $\Pr(Y(X = x) = y)$ can be constructed in the usual ways from the observed values of X and Y .

In a well-run randomized controlled experiment, strong ignorability of treatment assignment and SUTVA are likely to hold because of the design of the experiment. However, in situations other than the ideal experiment the assumptions necessary for the equality in Equation 1 to hold are unlikely to be satisfied. Thus, causal inferences made directly from the observed cell frequencies in 2×2 tables of non-experimental data are unlikely to be accurate.

However, in most applications it is generally believed that some (potentially unobservable) collection of variables \mathbf{U} exists such that treatment assignment is conditionally ignorable given \mathbf{U} .⁸ If one is also willing to assume that SUTVA holds, one can write the post-intervention distribution as:

$$\Pr(Y(X = x) = y) = \int \Pr(Y = y|X = x, \mathbf{U} = \mathbf{u})dP_{\mathbf{U}} \quad (2)$$

where $P_{\mathbf{U}}$ is the distribution of \mathbf{U} (Rosenbaum and Rubin, 1983b; Pearl, 1995). Note that \mathbf{U} may well be multivariate, continuous, and may well be unobservable. It is common to say that \mathbf{U} contains a set of confounding variables. Similarly, one often says that confounding is present when one needs to adjust for \mathbf{U} as in Equation 2. Heuristically, one can think of \mathbf{U} as containing variables that are common causes of X and Y .

If one is willing to assume that \mathbf{U} is rich enough so that the potential outcomes are a deterministic function of x_i and \mathbf{u}_i a tremendous simplification becomes possible.⁹ While \mathbf{U} may be extremely complicated, the binary nature of treatment and outcome implies that the domain of \mathbf{U} can be partitioned into four equivalence classes depending on the pattern of potential outcomes associated

⁸Formally, $[Y(X = 0), Y(X = 1)] \perp\!\!\!\perp X|\mathbf{U}$. Here strong ignorability of treatment assignment holds within each level of \mathbf{U} .

⁹Such an assumption is actually not a strong assumption since, as we will see below, \mathbf{u} need not be observed for the main results that follow. Further, most counterfactual causal models (see Rubin (1974, 1978); Pearl (2000)) explicitly assume deterministic potential outcomes at a conceptual level.

$Y_i(X_i = 0)$	$Y_i(X_i = 1)$	Z_i	
0	0	0	Never Succeed
0	1	1	Helped
1	0	2	Hurt
1	1	3	Always Succeed

Table 2: *Possible Patterns of Potential Outcomes and Coarsest General Confounding Variable.* A unit i for which $Z_i = 0$ has a value of U_i that causes it to always have $Y_i = 0$ regardless of the (counterfactual) value of X_i . We say these units are “never succeeders”. If $Z_i = 1$ we say that unit i is “helped” by treatment because its potential outcome under $X = 1$ is equal to 1 (success) while its potential outcome under $X = 0$ is 0 (failure). If unit i has $Z_i = 2$ we say that i is “hurt” by treatment because its potential outcome under $X = 1$ is equal to 0 (failure) while its potential outcome under $X = 0$ is 1 (success). Finally, if $Z_i = 3$ we say that i is an “always succeeder” because its value of U_i is such that Y_i will always equal 1 regardless of the (counterfactual) value of X_i .

with each point in the domain of \mathbf{U} (Angrist et al., 1996; Balke and Pearl, 1997; Chickering and Pearl, 1997). We introduce a new categorical variable Z_i that labels these equivalence classes. The values of Z_i along with the associated patterns of potential outcomes are presented in Table 2. Since Z is defined in terms of the value of the potential outcome pairs it is clear that conditional ignorability holds given Z . We assume that (X_i, Y_i, Z_i) for all i are independent replicates from some joint distribution P_{XYZ} . If P_{XYZ} is known, one could write the post-intervention distribution as

$$\Pr(Y(X = x) = y) = \sum_{z=0}^3 \Pr(Y = y|X = x, Z = z) \Pr(Z = z).$$

where the probabilities on the right-hand-side of the equation above can be calculated directly from P_{XYZ} . Similarly, if a sample was available from P_{XYZ} , one could use the sample values of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to consistently estimate P_{XYZ} and to then use this estimate to construct an estimate of the post-intervention distribution. Needless to say, it is extremely unlikely that applied researchers in the social sciences would find themselves in the situation where Z_i is observable for any i . Without data on Z it is impossible to consistently estimate P_{XYZ} . Nevertheless, there is some information about Z in observed (X, Y) data sampled from P_{XYZ} . The goal of this paper is to show how this information can be combined with subjective background knowledge to yield causal inferences from 2×2 and $2 \times 2 \times K$ tables even when the confounding variables in \mathbf{U} are not measured.

In what follows, we begin by specifying a probability model for the very unlikely situation where (x_i, y_i, z_i) are observed for all i . While this situation will not occur in practice, it does serve as a starting point to deal with the realistic situation of \mathbf{z} being completely unobserved. We deal with this more realistic situation in Section 2.2. Once again, the goal of this paper is to demonstrate how the situation in which \mathbf{z} is completely unobserved can still be analyzed so that accurate statements of subjective uncertainty regarding causal effects can still be made.

2.1 \mathbf{z} Completely Observed

The goal here is to make inferences about the form of P_{XYZ} and to then treat causal quantities as functionals of this distribution. We adopt a Bayesian approach. While Bayesian procedures have a number of advantages (Bernardo and Smith, 1994; Gelman et al., 2003; Gill, 2007), the main reason for taking a Bayesian approach in this paper is that it allows us to incorporate background knowledge about the (potentially unobserved) confounder Z in a principled fashion (Kadane and Wolfson, 1998; Western and Jackman, 1994; Gill and Walker, 2005). We begin by discussing the likelihood function and then discuss our choice of prior distribution along with the resulting posterior distribution.

2.1.1 Likelihood

Assuming that (X_i, Y_i, Z_i) for all i are independent replicates from some joint distribution P_{XYZ} we can write the likelihood as:

$$\begin{aligned}
p(\mathbf{x}, \mathbf{y}, \mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^n p(x_i, y_i, z_i | \boldsymbol{\theta}, \boldsymbol{\psi}) \\
&= \prod_{i=1}^n p(x_i, y_i | \boldsymbol{\theta}) p(z_i | x_i, y_i, \boldsymbol{\psi}) \\
&= \prod_{i=1}^n \theta_{00}^{\mathbb{I}(x_i=0, y_i=0)} \theta_{01}^{\mathbb{I}(x_i=0, y_i=1)} \theta_{10}^{\mathbb{I}(x_i=1, y_i=0)} \theta_{11}^{\mathbb{I}(x_i=1, y_i=1)} \times \\
&\quad \psi_{00}^{\mathbb{I}(x_i=0, y_i=0, z_i=1)} (1 - \psi_{00})^{\mathbb{I}(x_i=0, y_i=0, z_i=0)} \times \\
&\quad \psi_{01}^{\mathbb{I}(x_i=0, y_i=1, z_i=3)} (1 - \psi_{01})^{\mathbb{I}(x_i=0, y_i=1, z_i=2)} \times \\
&\quad \psi_{10}^{\mathbb{I}(x_i=1, y_i=0, z_i=2)} (1 - \psi_{10})^{\mathbb{I}(x_i=1, y_i=0, z_i=0)} \times \\
&\quad \psi_{11}^{\mathbb{I}(x_i=1, y_i=1, z_i=3)} (1 - \psi_{11})^{\mathbb{I}(x_i=1, y_i=1, z_i=1)} \\
&= \theta_{00}^{C_{00+}} \theta_{01}^{C_{01+}} \theta_{10}^{C_{10+}} \theta_{11}^{C_{11+}} \psi_{00}^{C_{001}} (1 - \psi_{00})^{C_{000}} \psi_{01}^{C_{013}} (1 - \psi_{01})^{C_{012}} \times \\
&\quad \psi_{10}^{C_{102}} (1 - \psi_{10})^{C_{100}} \psi_{11}^{C_{113}} (1 - \psi_{11})^{C_{111}} \tag{3}
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $C_{xyz} = \sum_{i=1}^n \mathbb{I}(x_i = x, y_i = y, z_i = z)$, $C_{xy+} = \sum_{i=1}^n \mathbb{I}(x_i = x, y_i = y)$, $\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11} \geq 0$, $\theta_{00} + \theta_{01} + \theta_{10} + \theta_{11} = 1$, and $\psi_{00}, \psi_{01}, \psi_{10}, \psi_{11} \in [0, 1]$. Note that C_{xyz} is the number of cases in cell (x, y, z) of the $2 \times 2 \times 4$ contingency table for (X, Y, Z) and that C_{xy+} is the number of cases in cell (x, y) of the 2×2 marginal table for (X, Y) .

We emphasize that, after assuming that X and Y are binary and (X_i, Y_i, Z_i) are independent replicates from P_{XYZ} , we have made no assumptions about the form of P_{XYZ} . Any logically consistent distribution P_{XYZ} and any form of confounding among independent replicates can be captured by particular values of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. This probability model for (X, Y, Z) is thus completely general and nonparametric.

While this model for (X, Y, Z) might seem to contain a large number of parameters that are difficult to interpret it is actually the case that there are a relatively small number of free parameters

Parameter	Probability	Interpretation
θ_{xy}	$\Pr(X_i = x, Y_i = y)$	Probability X_i is equal to x and Y_i is equal to y
ψ_{00}	$\Pr(Z_i = 1 X_i = 0, Y_i = 0)$	Probability i would be helped by treatment given i not treated and i failed
$1 - \psi_{00}$	$\Pr(Z_i = 0 X_i = 0, Y_i = 0)$	Probability i would never succeed given i not treated and i failed
ψ_{01}	$\Pr(Z_i = 3 X_i = 0, Y_i = 1)$	Probability i would always succeed given i not treated and i succeeded
$1 - \psi_{01}$	$\Pr(Z_i = 2 X_i = 0, Y_i = 1)$	Probability i would be hurt by treatment given i not treated and i succeeded
ψ_{10}	$\Pr(Z_i = 2 X_i = 1, Y_i = 0)$	Probability i was hurt by treatment given i treated and i failed
$1 - \psi_{10}$	$\Pr(Z_i = 0 X_i = 1, Y_i = 0)$	Probability i would never succeed given i treated and i failed
ψ_{11}	$\Pr(Z_i = 3 X_i = 1, Y_i = 1)$	Probability i would always succeed given i treated and i succeeded
$1 - \psi_{11}$	$\Pr(Z_i = 1 X_i = 1, Y_i = 1)$	Probability i was helped by treatment given i treated and i succeeded

Table 3: *Interpretation of Parameters in the Model for (X, Y, Z) .* The i indices denote a randomly selected unit.

that are easily interpretable in terms of conditional and marginal probabilities. Table 3 provides a summary of the key parameters and their intuitive meanings.

Here we see that there are two key sets of parameters θ_{xy} and ψ_{xy} for $x = 0, 1$ and $y = 0, 1$. The θ parameters govern a multinomial distribution for the distribution of (X, Y) after Z has been marginalized out of P_{XYZ} . For instance, θ_{01} is thus the probability that $X = 0$ and $Y = 1$ averaged over all 4 levels of Z . The ψ parameters govern the conditional distribution of Z given X and Y . Note that because of the definition of Z (see Table 2) only 2 values of Z are logically possible given any admissible (X, Y) pair. For instance, if $X = 0$ and $Y = 1$ then Z must be equal to either 2 or 3. The distribution of Z_i given $X_i = x$ and $Y_i = y$ is thus Bernoulli with parameter ψ_{xy} . Keeping with our example, ψ_{01} gives the probability that $Z_i = 3$ given $X_i = 0$ and $Y_i = 1$ while $(1 - \psi_{01})$ gives the probability that $Z_i = 2$ given $X_i = 0$ and $Y_i = 1$. The other conditional distributions for Z given $X = x$ and $Y = y$ are similarly parameterized.

2.1.2 Prior and Posterior Distributions

Bayesian inference centers on the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ given the observed data.

The posterior distribution is given (up to proportionality) by:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{y}, \mathbf{x}, \mathbf{z}) \propto p(\mathbf{y}, \mathbf{x}, \mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi})$$

The likelihood $p(\mathbf{y}, \mathbf{x}, \mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\psi})$ was given in the previous section. What is still needed is a specification of the prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\psi})$. A natural choice for the joint prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is to assume that $\boldsymbol{\theta}$, $\psi_{00}, \psi_{01}, \psi_{10}$, and ψ_{11} are mutually independent a priori and that $\boldsymbol{\theta} \sim \text{Dirichlet}(a_{00}, a_{01}, a_{10}, a_{11})$, $\psi_{xy} \sim \text{Beta}(b_{xy}, c_{xy})$, for $x = 0, 1$ and $y = 0, 1$. This is the conjugate prior distribution for this model. As we will see below, this prior specification will allow us to think of the hyper-parameters a_{xy}, b_{xy} , and c_{xy} for $x = 0, 1$ and $y = 0, 1$ as additional “pseudo-observations”. This makes the prior distributions more easily interpretable, which is quite important for the current application where inferences will typically be quite dependent on the prior.

Using the prior specification discussed above, we can write the posterior density (up to proportionality) as:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}, \mathbf{y}, \mathbf{z}) \propto & \theta_{00}^{C_{00+} + a_{00} - 1} \theta_{01}^{C_{01+} + a_{01} - 1} \theta_{10}^{C_{10+} + a_{10} - 1} \theta_{11}^{C_{11+} + a_{11} - 1} \times \\ & \psi_{00}^{C_{001} + b_{00} - 1} (1 - \psi_{00})^{C_{000} + c_{00} - 1} \psi_{01}^{C_{013} + b_{01} - 1} (1 - \psi_{01})^{C_{012} + c_{01} - 1} \times \\ & \psi_{10}^{C_{102} + b_{10} - 1} (1 - \psi_{10})^{C_{100} + c_{10} - 1} \psi_{11}^{C_{113} + b_{11} - 1} (1 - \psi_{11})^{C_{111} + c_{11} - 1} \end{aligned} \quad (4)$$

This is just the product of a Dirichlet distribution and four beta distributions. Note the similarity of the posterior in Equation 4 to the likelihood in Equation 3. Specifically, note that our prior for $\boldsymbol{\theta}$ is serving to add an additional $(a_{00} - 1)$ ($X = 0, Y = 0$) pairs, $(a_{01} - 1)$ ($X = 0, Y = 1$) pairs, etc. to the dataset. Similarly, our prior for $\boldsymbol{\psi}$ is adding an additional $(b_{00} - 1)$ ($X = 0, Y = 0, Z = 1$) triples, $(c_{00} - 1)$ ($X = 0, Y = 0, Z = 0$) triples, etc. to the dataset.

2.2 z Completely Unobserved

We now analyze the case in which the confounder Z is never observed in the sample data. Given the definition of Z , we expect that all applications will fall into this category. While inference in this situation relies critically on untestable assumptions about the relationship between Z and (X, Y) it is the case that there is some information in the observed C_{xy+} counts as to the distribution of Z in the population. Further, in many cases, there will be enough scientific background knowledge such that ψ can be given a defensible subjective prior distribution.

2.2.1 Likelihood

Let \mathcal{Z}_i denote the set of possible values Z_i could take given the observed data on unit i . More formally,

$$\mathcal{Z}_i = \begin{cases} \{0, 1\} & \text{if } x_i = 0, y_i = 0 \\ \{2, 3\} & \text{if } x_i = 0, y_i = 1 \\ \{0, 2\} & \text{if } x_i = 1, y_i = 0 \\ \{1, 3\} & \text{if } x_i = 1, y_i = 1 \end{cases}$$

We can then write the likelihood as:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^n \sum_{z_i \in \mathcal{Z}_i} p(x_i, y_i, z_i | \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^n p(x_i, y_i | \boldsymbol{\theta}) \left\{ \sum_{z_i \in \mathcal{Z}_i} p(z_i | x_i, y_i, \boldsymbol{\psi}) \right\} \\ &= \prod_{i=1}^n \theta_{00}^{\mathbb{I}(x_i=0, y_i=0)} \theta_{01}^{\mathbb{I}(x_i=0, y_i=1)} \theta_{10}^{\mathbb{I}(x_i=1, y_i=0)} \theta_{11}^{\mathbb{I}(x_i=1, y_i=1)} \times \\ &\quad \left\{ \sum_{z_i \in \mathcal{Z}_i} \psi_{00}^{\mathbb{I}(x_i=0, y_i=0, z_i=1)} (1 - \psi_{00})^{\mathbb{I}(x_i=0, y_i=0, z_i=0)} \times \right. \\ &\quad \psi_{01}^{\mathbb{I}(x_i=0, y_i=1, z_i=3)} (1 - \psi_{01})^{\mathbb{I}(x_i=0, y_i=1, z_i=2)} \times \\ &\quad \psi_{10}^{\mathbb{I}(x_i=1, y_i=0, z_i=2)} (1 - \psi_{10})^{\mathbb{I}(x_i=1, y_i=0, z_i=0)} \times \\ &\quad \left. \psi_{11}^{\mathbb{I}(x_i=1, y_i=1, z_i=3)} (1 - \psi_{11})^{\mathbb{I}(x_i=1, y_i=1, z_i=1)} \right\} \\ &= \theta_{00}^{C_{00+}} \theta_{01}^{C_{00+}} \theta_{10}^{C_{10+}} \theta_{11}^{C_{11+}} \end{aligned} \tag{5}$$

where θ_{xy} and C_{xy+} are as before for $x = 0, 1$ and $y = 0, 1$. Note that the complete missingness of Z has caused the term consisting of conditional probabilities of Z given X and Y to drop completely out of the likelihood. As we might expect, the observed data provide no information about ψ . Note, however, that this does not mean that there is no information in the observed data about Z . Because the domain of Z_i is cut in half from 4 possible values to 2 possible values conditional on x_i and y_i , there is some information in the data about the distribution of Z . This is important because, as we will see later, many causal quantities of interest can be written as functions of certain marginal probabilities of Z .

2.2.2 Prior and Posterior Distributions

Again, the prior discussed in Section 2.1.2 remains a natural choice for this situation. Combining this prior with the likelihood in Equation 5 gives us the following posterior density (up to proportionality):

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}, \mathbf{y}) \propto & \theta_{00}^{C_{00+} + a_{00} - 1} \theta_{01}^{C_{01+} + a_{01} - 1} \theta_{10}^{C_{10+} + a_{10} - 1} \theta_{11}^{C_{11+} + a_{11} - 1} \times \\
& \psi_{00}^{b_{00} - 1} (1 - \psi_{00})^{c_{00} - 1} \psi_{01}^{b_{01} - 1} (1 - \psi_{01})^{c_{01} - 1} \times \\
& \psi_{10}^{b_{10} - 1} (1 - \psi_{10})^{c_{10} - 1} \psi_{11}^{b_{11} - 1} (1 - \psi_{11})^{c_{11} - 1}
\end{aligned} \tag{6}$$

Note that the only information about $\boldsymbol{\psi}$ is coming from the prior distribution. This implies that inferences that depend on $\boldsymbol{\psi}$ will typically be sensitive to one's choice of prior for $\boldsymbol{\psi}$.

3 Causal Quantities of Interest

The purpose of this section is to show how the causal quantities discussed at the beginning of Section 2 can be calculated from $\boldsymbol{\theta}$ and, in some cases, $\boldsymbol{\psi}$. Because the form of these causal quantities will depend on what (conditional) ignorability assumptions one is willing to make, we need a terminology to distinguish these different forms of the quantities of interest. Following Holland (1986) we term quantities that would be true causal quantities under strong ignorability

prima facie quantities. These are the quantities that would be calculated directly from the the observed data under an assumption of unconfoundedness. In most non-experimental situations the *prima facie* quantities will not be true causal quantities. The other types of quantities we are interested in here are those that are generally valid causal quantities but depend on the typically unidentified parameter ψ . We term these quantities *sensitivity analysis* quantities. While these quantities will be genuine causal quantities as long as SUTVA holds, the fact that the confounding variable Z is not typically observed means that inference will be sensitive to prior beliefs about the conditional distribution of Z in the population.

3.1 Post Intervention Distribution

As noted in Section 2, the starting point for the derivation of all causal quantities of interest is the post-intervention distribution. This is the distribution over the potential outcomes given a hypothetical manipulation of X applied to the entire population of units. We can write the *prima facie* post intervention distribution as:

$$\begin{aligned} \Pr_p(Y(X = 0) = 0) &= \Pr(Y = 0|X = 0) \\ &= \frac{\theta_{00}}{\theta_{00} + \theta_{01}} \\ \Pr_p(Y(X = 0) = 1) &= \Pr(Y = 1|X = 0) \\ &= \frac{\theta_{01}}{\theta_{00} + \theta_{01}} \\ \Pr_p(Y(X = 1) = 0) &= \Pr(Y = 0|X = 1) \\ &= \frac{\theta_{10}}{\theta_{10} + \theta_{11}} \\ \Pr_p(Y(X = 1) = 1) &= \Pr(Y = 1|X = 1) \\ &= \frac{\theta_{11}}{\theta_{10} + \theta_{11}}. \end{aligned}$$

As we might expect, the *prima facie* post-intervention distribution does not depend on ψ and thus can be consistently estimated regardless of whether or not Z is observed. The *prima facie* post intervention distribution will be the true post intervention distribution if there is no confounding—

i.e., treatment assignment is strongly ignorable. All other prima facie causal quantities can be derived from the prima facie post-intervention.

The sensitivity analysis post intervention distribution is given by:

$$\begin{aligned}
\Pr_s(Y(X = 0) = 0) &= \sum_{z=0}^3 \Pr(Y = 0|X = 0, Z = z) \Pr(Z = z) \\
&= \Pr(Z = 0) + \Pr(Z = 1) \\
&= \theta_{10}(1 - \psi_{10}) + \theta_{11}(1 - \psi_{11}) + \theta_{00} \\
\Pr_s(Y(X = 0) = 1) &= \sum_{z=0}^3 \Pr(Y = 1|X = 0, Z = z) \Pr(Z = z) \\
&= \Pr(Z = 2) + \Pr(Z = 3) \\
&= \theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01} \\
\Pr_s(Y(X = 1) = 0) &= \sum_{z=0}^3 \Pr(Y = 0|X = 1, Z = z) \Pr(Z = z) \\
&= \Pr(Z = 0) + \Pr(Z = 2) \\
&= \theta_{00}(1 - \psi_{00}) + \theta_{01}(1 - \psi_{01}) + \theta_{10} \\
\Pr_s(Y(X = 1) = 1) &= \sum_{z=0}^3 \Pr(Y = 1|X = 1, Z = z) \Pr(Z = z) \\
&= \Pr(Z = 1) + \Pr(Z = 3) \\
&= \theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}
\end{aligned}$$

Because conditional ignorability holds according to the definition of Z , if ψ were known the sensitivity analysis post-intervention distribution would yield the true post intervention distribution. Of course, ψ is never known (and typically not identified) so the sensitivity analysis post-intervention distribution will depend on one's prior beliefs about ψ . Note, however, that this sensitivity to ψ does not imply that there is *no* information in the observed data about the sensitivity analysis post-intervention distribution and functionals thereof. We explore the the extent to which the observed data alone are informative about causal quantities below.

3.2 Average Treatment Effects

The average treatment effect is what most social scientists think of as “the” causal effect of X on Y . As noted above, it describes the expected difference of $Y(X = 1)$ and $Y(X = 0)$ for a unit randomly chosen from the study population. While the focus in this subsection is simply on the average treatment effect or ATE , it is also possible to define average treatment effects within just the treated group (ATT) or just the control group (ATC) and calculate bounds and sensitivity analysis distributions for these estimands as well. We do not do that here for reasons of space. The software that accompanies this article does allow the user to calculate ATT and ATC along with ATE .

3.2.1 Prima Facie Average Treatment Effect

The prima facie average treatment effect is defined as:

$$\begin{aligned} ATE_p &= \Pr_p(Y(X = 1) = 1) - \Pr_p(Y(X = 0) = 1) \\ &= \frac{\theta_{11}}{\theta_{10} + \theta_{11}} - \frac{\theta_{01}}{\theta_{00} + \theta_{01}} \end{aligned}$$

If treatment assignment is strongly ignorable and SUTVA holds it is easy to show that ATE_p equals the true average treatment effect in the population.

3.2.2 Sensitivity Analysis Average Treatment Effect

The sensitivity analysis average treatment effect is defined as:

$$\begin{aligned} ATE_s &= \Pr_s(Y(X = 1) = 1) - \Pr_s(Y(X = 0) = 1) \\ &= (\theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}) - (\theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01}) \end{aligned} \tag{7}$$

Because of the definition of Z , ATE_s will always equal the true average treatment effect as long as SUTVA holds.

3.2.3 Large Sample Nonparametric Bounds for the Average Treatment Effect

Following Manski (1990) we can derive nonparametric bounds for the average treatment effect that will contain the true average treatment effect with probability 1 as sample size goes to infinity. Inspection of Equation 7 reveals that the minimum value of ATE_s will occur when $\psi_{00} = 0, \psi_{01} = 0, \psi_{10} = 1$, and $\psi_{11} = 1$. Similarly, the maximum value of ATE_s will occur when $\psi_{00} = 1, \psi_{01} = 1, \psi_{10} = 0$, and $\psi_{11} = 0$. Substituting these values into the expression for ATE_s and recognizing that $ATE_s = ATE$ we see that:

$$ATE \in [-(\theta_{10} + \theta_{01}), (\theta_{00} + \theta_{11})]$$

Substituting the MLEs for $\theta_{00}, \theta_{01}, \theta_{10}$, and θ_{11} we see that (in a slight abuse of notation)

$$\lim_{n \rightarrow \infty} \Pr \left(\frac{-C_{01+} - C_{10+}}{n} \leq ATE \leq \frac{C_{00+} + C_{11+}}{n} \right) = 1$$

where it is understood that the C_{xy+} counts also depend on n . Note that this interval will always include 0. Further, as Manski (1990) has shown, and as is easy to see here since $\sum_x \sum_y \theta_{xy} = 1$, the width of this interval will always be 1.

3.3 (Log) Relative Risk

The relative risk is the ratio of two post-intervention probabilities. It tells how many times more likely a positive outcome is under treatment than under control. Relative risks are often of interest in situations where the outcome variable describes a rare event. Here the ATE may be quite small in absolute value while the relative risk is large. See King and Zeng (2002) for a discussion of the concept of relative risk and related concepts.

3.3.1 Prima Facie (Log) Relative Risk

The prima facie relative risk is defined as:

$$\begin{aligned} RR_p &= \frac{\Pr_p(Y(X=1)=1)}{\Pr_p(Y(X=0)=1)} \\ &= \frac{\theta_{11}}{\theta_{10} + \theta_{11}} \bigg/ \frac{\theta_{01}}{\theta_{00} + \theta_{01}} \end{aligned}$$

It will often be convenient to report the log of RR_p which is referred to as the prima facie log relative risk.

3.3.2 Sensitivity Analysis (Log) Relative Risk

The sensitivity analysis relative risk is defined as:

$$\begin{aligned} RR_s &= \frac{\Pr_s(Y(X=1)=1)}{\Pr_s(Y(X=0)=1)} \\ &= \frac{\theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}}{\theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01}} \end{aligned} \quad (8)$$

3.3.3 Large Sample Nonparametric Bounds for (Log) Relative Risks

Looking at Equation 8 we see that the minimum value of RR_s will occur when $\psi_{00} = 0, \psi_{01} = 0, \psi_{10} = 1$, and $\psi_{11} = 1$. Similarly, the maximum value of RR_s will occur when $\psi_{00} = 1, \psi_{01} = 1, \psi_{10} = 0$, and $\psi_{11} = 0$. Substituting these values into the expression for RR_s and recognizing that $RR_s = RR$ we see that:

$$RR \in \left[\frac{\theta_{11}}{\theta_{01} + \theta_{10} + \theta_{11}}, \quad \frac{\theta_{00}}{\theta_{01}} + \frac{\theta_{11}}{\theta_{01}} + 1 \right]$$

Note that this interval will always include 1.

Substituting the MLEs for $\theta_{00}, \theta_{01}, \theta_{10}$, and θ_{11} we see that (in an abuse of notation)

$$\lim_{n \rightarrow \infty} \Pr \left(\frac{C_{11+}}{C_{01+} + C_{10+} + C_{11+}} \leq RR \leq \frac{C_{00+}}{C_{01+}} + \frac{C_{11+}}{C_{01+}} + 1 \right) = 1$$

where it is understood that the C_{xy+} counts depend on n . The large sample bounds for the log relative risk can be obtained by logging the endpoints of the interval above. Unlike the bounds for the *ATE* which must always have width 1, the width of the bounding interval for the *RR* is only constrained to be greater than or equal to 1.

4 Bayesian Inference For Causal Effects

As we have seen in Sections 3.2.3 and 3.3.3, there is some information in just the marginal counts C_{xy+} about causal quantities such as the average treatment effect and the relative risk.

However, we have also seen that these bounds, taken on their own, cannot rule out the null values of 0 and 1 for the *ATE* and *RR* respectively. Further, the large sample bounds do not allow us to make statements about the probability that the quantity of interest is within some subset of the bounding interval. A third problem is that while the bounds are generally valid as sample size gets large it is less clear what one should believe about the likely value of the quantity of interest in small samples given just the information in the bounds.

A principled method to address all of these issues is to adopt a Bayesian approach and base inference on the posterior density in Equation 6. Taking such an approach requires one to specify a prior distribution for (θ, ψ) . In Section 2.1.2 we argued that independent Dirichlet and Beta distributions made sense in terms of interpretability. In the remainder of this section we discuss how the parameters governing these prior distributions can be chosen and how one can easily summarize the resulting posterior distribution to make inferences about causal quantities of interest such as *ATE* and *RR*.

4.1 Choosing a Prior Distribution

It is worth emphasizing that, unlike Bayesian inference for models which are point identified, the impact of the choice of prior for ψ on the posterior distribution for ψ and functionals of that posterior distribution will not diminish as n gets large if Z is completely unobserved. In fact, since no new information about ψ is arriving as n gets large, the marginal posterior for ψ will always be equal to the prior for ψ . Clearly one needs to be quite careful in determining the prior for ψ . Specifically, one should either be able to justify a particular choice of prior by an appeal to substantive background knowledge and/or perform a prior sensitivity analysis in which multiple reasonable priors—including some distributions that are based on the most extreme but still defensible assumptions that critical subject matter experts would entertain—are tried and the associated results are reported.

As is summarized in Tables 2 and 3 each ψ_{xy} represents the conditional probability of one of the two possible configurations of potential outcomes among units in which we observe $X = x$

and $Y = y$. Thus the $Beta(b_{xy}, c_{xy})$ prior for ψ_{xy} can be thought of as a statement of belief that $b_{xy} - 1$ of the C_{xy+} units have one potential outcome profile while $c_{xy} - 1$ of the C_{xy+} units have the other possible potential outcome profile. If $b_{xy} + c_{xy} = C_{xy+} + 2$ then the information in the prior is equivalent to the information that would be in the sample data in the ideal case in which the potential outcome patterns are observed for units with $X = x$ and $Y = y$. If $b_{xy} + c_{xy} < C_{xy+} + 2$ then there is less information in the prior than this ideal situation and if $b_{xy} + c_{xy} > C_{xy+} + 2$ then the prior is adding more information than one could ever get directly from the sample data.

To clarify the relationship between the b_{xy} and c_{xy} parameters and potential outcomes we will walk through the interpretation of ψ_{xy} , b_{xy} , and c_{xy} for all x and y . An additional example of how this works will appear in the empirical example in Section 6.

Begin with the situation in which $X_i = 0$ and $Y_i = 0$. Here two potential outcome profiles are possible: $Z_i = 0$ (never succeed) and $Z_i = 1$ (helped). ψ_{00} is the conditional probability that $Z_i = 1$ (i would be helped by treatment) given $X_i = 0$ and $Y_i = 0$, while $1 - \psi_{00}$ is obviously the conditional probability that $Z_i = 0$ (i would never succeed) given $X_i = 0$ and $Y_i = 0$. b_{00} is the number of pseudo $Z = 1$ (helped) observations + 1 and c_{00} is the number of pseudo $Z = 0$ (never succeed) observations + 1. If our background knowledge suggests that units for which we observe $X = 0$ and $Y = 0$ are unlikely to respond to treatment we would set $c_{00} > b_{00}$. This implies that $\Pr(Z_i = \text{never succeed} | X_i = 0, Y_i = 0) > \Pr(Z_i = \text{helped} | X_i = 0, Y_i = 0)$. On the other hand, if we believe that these units are more likely than not to respond to treatment we would set $b_{00} > c_{00}$. This would imply $\Pr(Z_i = \text{helped} | X_i = 0, Y_i = 0) > \Pr(Z_i = \text{never succeed} | X_i = 0, Y_i = 0)$. The absolute magnitude of b_{00} and c_{00} determines how sure we are of the potential outcome distribution within the $X = 0, Y = 0$ group.

Next consider the situation in which $X_i = 0$ and $Y_i = 1$. Again, two potential outcome profiles are possible: $Z_i = 2$ (hurt) and $Z_i = 3$ (always succeed). ψ_{01} is the conditional probability that $Z_i = 3$ (i would always succeed) given $X_i = 0$ and $Y_i = 1$, while $1 - \psi_{01}$ is the conditional probability that $Z_i = 2$ (i would be hurt by treatment) given $X_i = 0$ and $Y_i = 1$. b_{01} is the number of pseudo

$Z = 3$ (always succeed) observations + 1 and c_{01} is the number of pseudo $Z = 2$ (hurt) observations + 1. If our background knowledge suggests that units for which we observe $X = 0$ and $Y = 1$ are unlikely to respond to treatment we would set $b_{01} > c_{01}$. On the other hand, if we believe that these units are more likely than not to respond (negatively) to treatment we would set $c_{01} > b_{01}$. Again, the absolute magnitude of b_{01} and c_{01} determines how sure we are of the potential outcome distribution within the $X = 0, Y = 1$ group.

Within the $X_i = 1$ and $Y_i = 0$ group the two potential outcome profiles are $Z_i = 0$ (never succeed) and $Z_i = 2$ (hurt) . ψ_{10} is the conditional probability of randomly selecting a unit for which $Z_i = 2$ (i is hurt by treatment) from the $X_i = 1$ and $Y_i = 0$ group. b_{10} is the number of pseudo $Z = 2$ (hurt) observations + 1 and c_{10} is the number of pseudo $Z = 0$ (never succeed) observations + 1. Setting $b_{10} > c_{10}$ would be consistent with a prior belief that more subjects tend to respond negatively to treatment within this group, while setting $c_{10} > b_{10}$ would be consistent with a belief that treatment is more likely than not to have no effect on units within this group.

Finally, within the $X_i = 1$ and $Y_i = 1$ group we see that the two potential outcome profiles are $Z_i = 1$ (helped) and $Z_i = 3$ (always succeed). ψ_{11} is the conditional probability of seeing a $Z_i = 3$ (always succeed) observation within this group. b_{11} is the number of pseudo $Z = 3$ (always succeed) observations + 1 and c_{11} is the number of pseudo $Z = 1$ (helped) observations + 1. If one thinks that units within this group are, on average, likely to respond positively to treatment one would set $c_{11} > b_{11}$. If non-responsiveness is hypothesized one would set $b_{11} > c_{11}$.

Now that we have discussed the basic interpretive issues of setting a subjective prior for ψ we can move on to discuss some more general strategies that be used to operationalize such a prior. In some settings it is reasonable to assume that the treatment has a monotonic effect on outcomes for most units under study. Put more simply, under an assumption of a generally positive (negative) monotonic effect, the treatment may have either a positive (negative) effect or no effect on most units but it is unlikely to have a negative (positive) effect on any but a small fraction of units. If one believes that a generally positive monotonic treatment effect is

reasonable then one could set $b_{01} \gg c_{01}$ and $b_{10} \ll c_{10}$. Conversely, one could set $b_{00} \ll c_{00}$ and $b_{11} \gg c_{11}$ to operationalize a generally negative monotonic treatment effect. In some sense, Manski’s (2003) monotone treatment response assumption is a limiting case of this approach in which under a positive (negative) monotone treatment response there are absolutely no units that are hurt (helped) by treatment. Just as in Manski’s bounds analysis, the assumption of a generally a monotone treatment effect can greatly narrow down the likely range of causal quantities of interest.

One of the major advantages of thinking of unmeasured confounding in terms of the conditional distribution of the potential outcomes given x_i and y_i is that it formulates the problem in a way that is relatively easy for practitioners to reason about while remaining completely general. In order to apply the methods presented here, a researcher only needs to be able to reason about eight easily understood quantities— only four of which are completely free parameters. This is true regardless of the form of unmeasured confounding in his or her study. Contrast this with approaches to sensitivity analysis that attempt to directly model the relationships between the confounding variable(s) and outcomes and the confounding variable(s) and treatment status (Schlesselman, 1978; Rosenbaum and Rubin, 1983a; Lin et al., 1998). Such approaches require the practitioner to make assumptions about a) the nature of the confounding variable (whether it is univariate/multivariate, discrete/continuous, etc.) as well as b) the association between this assumed confounder and outcomes and treatment status. As (Robins, 1999, p. 172) and (Brumback et al., 2004, p. 763) have argued, it is “essentially impossible” for researchers to reason about a) and that even if the nature of the confounding variables can be assumed it is difficult for most practitioners to reason about the necessary associations.

4.2 Posterior Inference

It is relatively easy to show that the posterior distributions for θ and ψ discussed in Sections 2.1 and 2.2 can all be sampled from using simple independent Monte Carlo sampling. Markov chain Monte Carlo is not necessary. Once a sample of (θ, ψ) is available from the posterior distribution

of interest it is a simple matter to plug the draws of these parameters into the formulas for the post-intervention distribution, average treatment effects, relative risk, etc. in order to obtain the posterior distribution over these causal quantities of interest. To produce a Monte Carlo sample of size m from the distribution with density given (up to proportionality) by Equation 6 we can use Algorithm 4.1.

```

Algorithm 4.1: POSTERIOR SAMPLING UNOBSERVED Z( $\mathbf{C}, \mathbf{a}, \mathbf{b}, \mathbf{c}, m$ )
for  $j \leftarrow 1$  to  $m$ 
do
   $\theta^{(j)} \leftarrow rdirichlet(C_{00+} + a_{00}, C_{01+} + a_{01}, C_{10+} + a_{10}, C_{11+} + a_{11})$ 
   $\psi_{00}^{(j)} \leftarrow rbeta(b_{00}, c_{00})$ 
   $\psi_{01}^{(j)} \leftarrow rbeta(b_{01}, c_{01})$ 
   $\psi_{10}^{(j)} \leftarrow rbeta(b_{10}, c_{10})$ 
   $\psi_{11}^{(j)} \leftarrow rbeta(b_{11}, c_{11})$ 
return  $(\{\theta^{(j)}\}_{j=1}^m, \{\psi_{00}^{(j)}\}_{j=1}^m, \{\psi_{01}^{(j)}\}_{j=1}^m, \{\psi_{10}^{(j)}\}_{j=1}^m, \{\psi_{11}^{(j)}\}_{j=1}^m)$ 

```

Here $rdirichlet(d, e, f, g)$ is a function that returns a pseudo-random draw from a $Dirichlet(d, e, f, g)$ distribution and $rbeta(d, e)$ is a function that returns a pseudo-random draw from a $Beta(d, e)$ distribution.

4.2.1 Inference for Causal Quantities of Interest

Once a sample $\{\theta^{(j)}, \psi^{(j)}\}_{j=1}^m$ from the posterior distribution of (θ, ψ) has been drawn it is quite easy to plug these draws into the formulas for causal quantities of interest given in Section 3 to obtain a sample from the posterior distribution of the causal quantity of interest.

For instance a sample from the posterior distribution of the prima facie average treatment effect ($\{ATE_p^{(j)}\}_{j=1}^m$) can be constructed by taking the j th sample to be

$$ATE_p^{(j)} = \frac{\theta_{11}^{(j)}}{\theta_{10}^{(j)} + \theta_{11}^{(j)}} - \frac{\theta_{01}^{(j)}}{\theta_{00}^{(j)} + \theta_{01}^{(j)}}$$

for $j = 1, \dots, m$.

Similarly, a sample from the posterior distribution of the sensitivity analysis average treatment

effect $(\{ATE_s^{(j)}\}_{j=1}^m)$ can be constructed by taking the j th sample to be

$$ATE_s^{(j)} = \left(\theta_{00}^{(j)} \psi_{00}^{(j)} + \theta_{01}^{(j)} \psi_{01}^{(j)} + \theta_{11}^{(j)} \right) - \left(\theta_{10}^{(j)} \psi_{10}^{(j)} + \theta_{11}^{(j)} \psi_{11}^{(j)} + \theta_{01}^{(j)} \right)$$

for $j = 1, \dots, m$. Samples from the posterior distributions of other causal quantities of interest follow analogously.

Once one has a sample from the posterior distribution of interest it is a simple matter to summarize the distribution by calculating density estimates, highest posterior density regions (the smallest region that contains a pre-specified amount of the posterior mass), the probability that a quantity of interest is greater than 0, etc. using the sampled parameter values. See Jackman (2000); King et al. (2000); Gelman et al. (2003) and Gill (2007) for discussions of how posterior samples can be summarized.

5 Ranges of ψ Consistent With a Given Prima Facie Post-Intervention Distribution: The Confounding Plot

While the subjective Bayesian approach outlined above takes beliefs about ψ as input and returns a subjective probability distribution over causal effects, we can also reverse this process and start with a particular prima facie post-intervention distribution and ask what values of ψ will result in a sensitivity analysis post-intervention distribution that is within some tolerance of the given prima facie post-intervention distribution.

Formally, for a given values of $\theta_{00}, \theta_{01}, \theta_{10}$ and θ_{11} we seek to find all values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} for which

$$|\Pr_p(Y(X=0)=0) - \Pr_s(Y(X=0)=0)| = \left| \frac{\theta_{00}}{\theta_{00} + \theta_{01}} - (\theta_{10}(1 - \psi_{10}) + \theta_{11}(1 - \psi_{11}) + \theta_{00}) \right| \leq \epsilon \quad (9)$$

and

$$|\Pr_p(Y(X=0)=1) - \Pr_s(Y(X=0)=1)| = \left| \frac{\theta_{01}}{\theta_{00} + \theta_{01}} - (\theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01}) \right| \leq \epsilon \quad (10)$$

and

$$|\Pr_p(Y(X = 1) = 0) - \Pr_s(Y(X = 1) = 0)| = \left| \frac{\theta_{10}}{\theta_{10} + \theta_{11}} - (\theta_{00}(1 - \psi_{00}) + \theta_{01}(1 - \psi_{01}) + \theta_{10}) \right| \leq \epsilon \quad (11)$$

and

$$|\Pr_p(Y(X = 1) = 1) - \Pr_s(Y(X = 1) = 1)| = \left| \frac{\theta_{11}}{\theta_{10} + \theta_{11}} - (\theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}) \right| \leq \epsilon \quad (12)$$

for some small positive ϵ . Note that Inequalities 9 and 10 only depend on ψ_{10} and ψ_{11} while Inequalities 11 and 12 only depend on ψ_{00} and ψ_{01} . It is thus possible to depict all values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} that satisfy Inequalities 9 - 12 with a pair of 2-dimensional plots— one of ψ_{10} and ψ_{11} and another of ψ_{00} and ψ_{01} . Figure 4 depicts how these plots look for particular value of $\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}$, and ϵ .

Because this method does not account for sampling variability it is most appropriate for situations in which all of the cells in the 2×2 table have a reasonably large number of observations. We note in passing the similarity of the plots above to the tomography plots of King (1997) that are useful for ecological inference. Indeed, the situation under consideration in this paper in which C_{xy+} are all fully observed but the joint C_{xyz} counts are not observed can be thought of as a particular type of ecological inference problem (Richardson, 2004).

6 Example: Jury Aversion and Voter Registration Revisited

In this section, the methods discussed previously will be applied to data from a study by Oliver and Wolfinger (1999) which looks at the extent to which perceiving jury lists to be constructed from voter registration lists causes a decrease in voter registration. As Oliver and Wolfinger note, there is a widespread belief among election officials and some scholars that the practice of drawing jury lists from voter registration records depresses voter registration and hence turnout. To estimate the extent to which aversion to jury duty depresses voter registration, Oliver and Wolfinger use data from the 1991 ANES Pilot Study (Miller et al., 1991). The simplest tabulation of the data used by Oliver and Wolfinger appears in Table 1.

If one were to assume that treatment assignment is strongly ignorable one could analyze this 2×2 table as if it were a randomized experiment. Doing so we see that the posterior mean of the prima facie average treatment effect is equal to -0.076 with associated 95% highest posterior density (HPD) region $[-0.134, -0.016]$ and the posterior mean of the prima facie log relative risk is -0.090 with 95% HPD region $[-0.159, -0.019]$. The posterior probability that the prima facie average treatment effect is less than 0 is 0.991. The same is true for the prima facie log relative risk.

These results would seem to suggest that within the portion of the population that has some knowledge of how jury lists are constructed (a little over half the ANES sample) a belief that voter registration records are used to build jury lists decreases voter registration by about 7 and a half percentage points. In terms of relative risk, those who do not think that jury lists are constructed from voter lists are about 9.4% more likely to register to vote than those who do think jury lists are built from voter lists. These results suggest a moderate negative effect that would be consistent with the jury aversion hypothesis. Nonetheless, since these results rely on the implausible assumption of ignorable treatment assignment one would be correct to regard these prima facie results with more than a little suspicion.

6.1 Bayesian Analysis of the 2×2 Table Under Non-Informative Priors

As an initial step toward determine what one should infer about the effect of jury aversion of voter registration we calculate the ATE_s and RR_s under three different non-informative priors for ψ . The first corresponds to a uniform prior over each element of ψ . Formally, $\psi_{xy} \sim \mathcal{Beta}(1, 1)$ for all x and y . The second prior corresponds to the Jeffreys prior $\psi_{xy} \sim \mathcal{Beta}(0.5, 0.5)$ for all x and y . We also examine results under the assumption that $\psi_{xy} \sim \mathcal{Beta}(0.001, 0.001)$.

Graphical depictions of these results along with the prima facie results and the large sample nonparametric bounds are presented in Figure 1. Here we see that the large sample nonparametric bounds for the average treatment effect are $ATE \in [-0.343, 0.657]$ and the bounds for the log relative risk are $\log(RR) \in [-0.434, 1.491]$. Assuming that there may be unmeasured confounding,

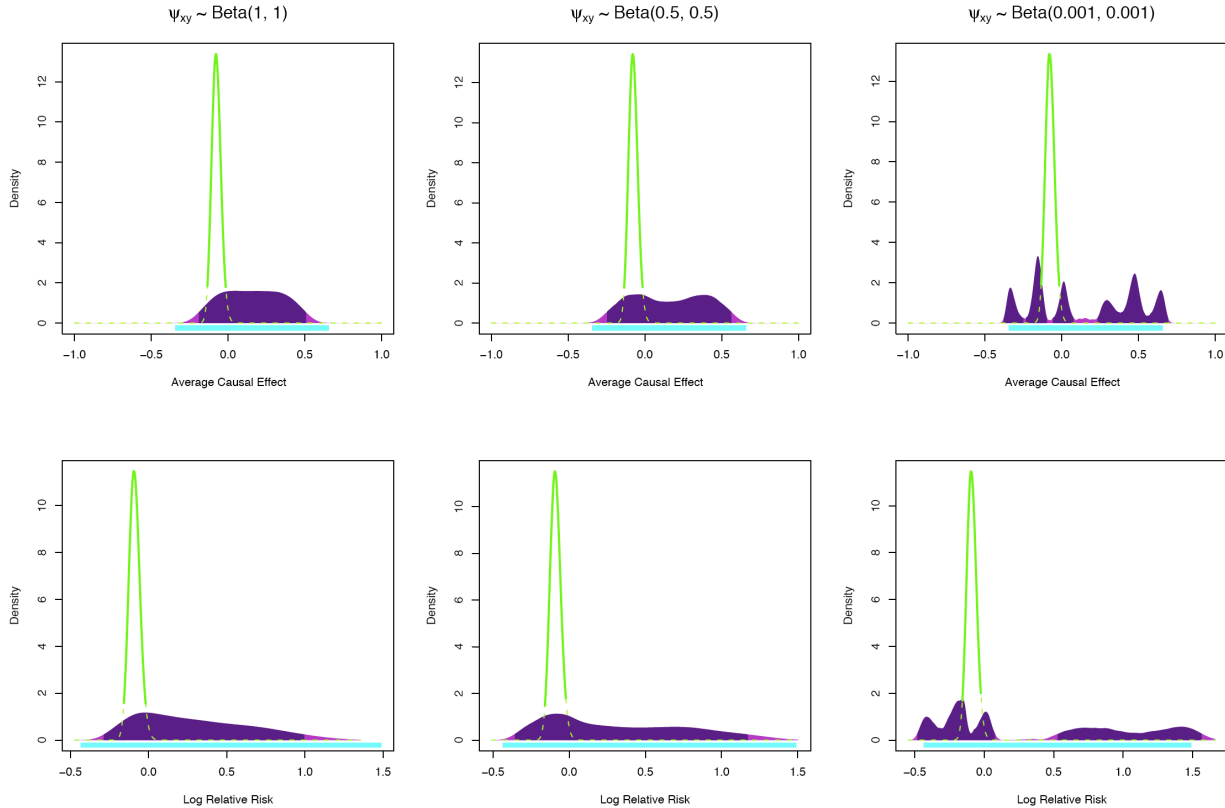


Figure 1: *Posterior Distributions of Prima Facie and Sensitivity Analysis Average Treatment Effects and Relative Risks for Data in Table 1 Under Non-Informative Priors.* The green line in each panel is the posterior density of the prima facie ATE or RR (solid line depicts 95% highest posterior density region). The purple polygons depict the posterior density of the sensitivity analysis ATE or RR under alternative priors for ψ (dark purple denotes the 95% highest posterior density region). The horizontal light blue line in each each figure denotes the region inside the large sample non-parametric bounds for the estimand. The leftmost panels are from the model that assumes uniform prior distributions for all ψ parameters. The center panels are from the model that assumes a $\mathcal{Beta}(0.5, 0.5)$ prior distributions for all ψ parameters. Finally, the rightmost panels are from the model that assumes $\mathcal{Beta}(0.001, 0.001)$ prior distributions for all ψ parameters.

but nothing at all about the form of that confounding, we should believe that the ATE and $\log(RR)$ are within the above intervals. However, we cannot say how likely these quantities are to be within any subregion of these intervals. To do that we need to specify a prior distribution for ψ .

Under the $\mathcal{Beta}(1, 1)$ prior we see that the posterior mean of the ATE_s is 0.157 with associated 95% HPD region $[-0.189, 0.506]$ and that the posterior mean of $\log(RR_s)$ is 0.294 with 95% HPD region $[-0.284, 0.990]$. Under the $\mathcal{Beta}(0.5, 0.5)$ prior the posterior mean of ATE_s is

0.157 with 95% HPD region $[-0.246, 0.562]$ while the posterior mean of $\log(RR_s)$ is 0.324 with 95% HPD region $[-0.356, 1.149]$. In both cases, there is little posterior mass near the endpoints of the large sample bounds. Something different happens when we adopt the $\text{Beta}(0.001, 0.001)$ prior. This prior puts most of the prior mass very near $\psi_{xy} = 0$ and $\psi_{xy} = 1$. As a result, the posterior distributions for ATE_s and $\log(RR_s)$ become multimodal. Further, because of the finite sample size, some non-negligible posterior mass is placed outside the large-sample bounds of both ATE and $\log(RR)$. Specifically, we see that the 95% HPD region for ATE_s is: $[-0.377, -0.258] \cup [-0.23, -0.097] \cup [-0.053, 0.076] \cup [0.235, 0.692]$ and the 95% HPD region for $\log(RR_s)$ is: $[-0.481, -0.096] \cup [-0.094, 0.076] \cup [0.538, 1.563]$. Indeed, one of the advantages of the approach taken in this paper is that it remains valid in finite samples of any size. On the whole, looking at these results under various non-informative prior distributions would suggest that there is not enough evidence to reject the null of $ATE = 0$ and $\log(RR) = 0$.

6.2 Bayesian Analysis of the 2×2 Table Under Informative Priors

While such an analysis under a series of non-informative prior distributions may seem compelling, one needs to remember that since no sample information about ψ is available, even these “non-informative” prior distributions are building in some assumptions that may be at odds with scientific knowledge or common sense. For instance, all three of the “non-informative” priors used above are symmetric around 0.5 and thus have the property that the prior density for ψ_{xy} is the same as the prior density for $1 - \psi_{xy}$. Further, since b_{xy} and c_{xy} were assumed to equal the same constant for all x and y it follows that the marginal prior distributions for all of the ψ_{xy} parameters are equivalent. This means, for instance, that these priors embody the assumption that the fraction of units hurt by treatment in the $(X = 1, Y = 0)$ group is equal to the fraction of units helped by treatment in the $(X = 0, Y = 0)$ group. This is an implausible assumption in most applications where there is even a hint of background knowledge— including the current application.

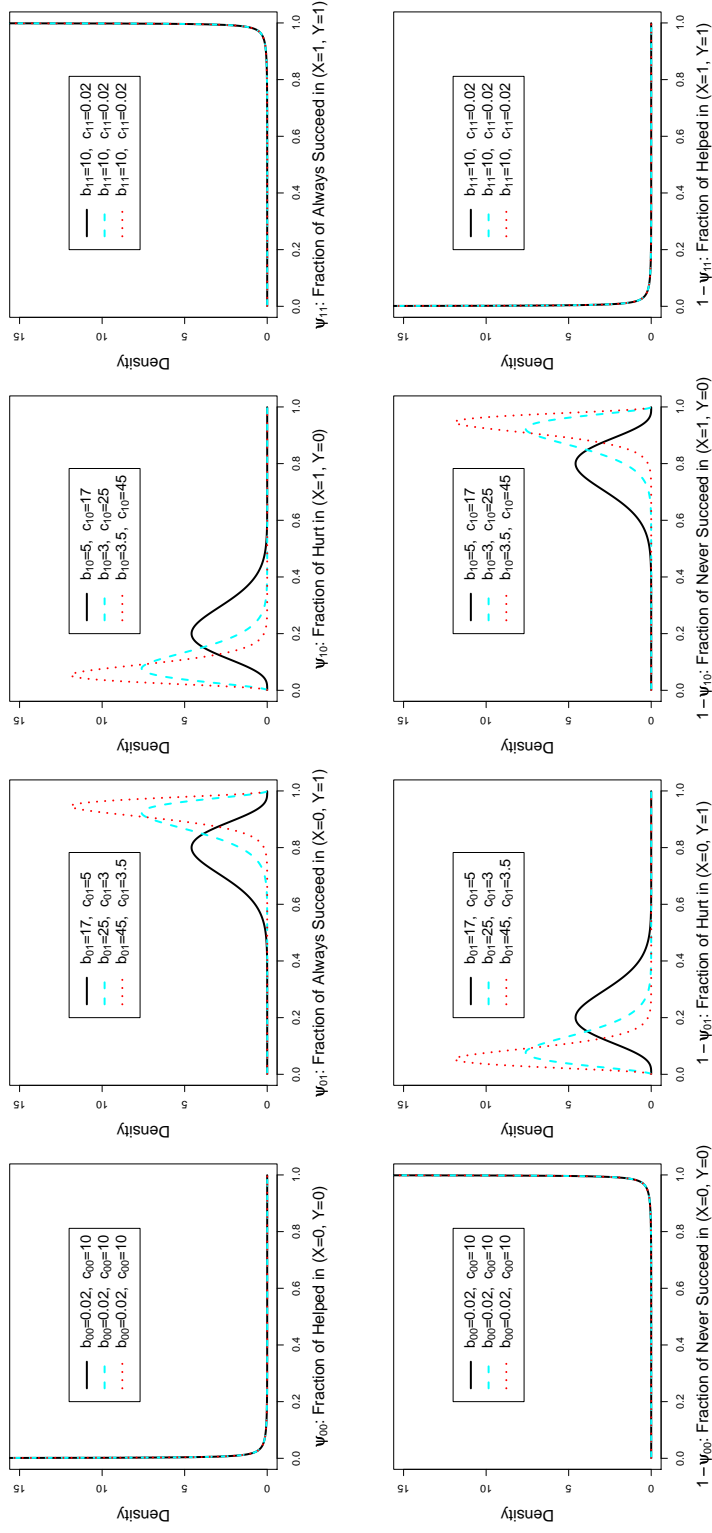


Figure 2: *Subjective Prior Densities Used in Analysis of Table 1.* The prior referred to as prior (a) is given by the solid black lines, the prior referred to as prior (b) is given by the dashed blue lines, and the prior referred to as prior (c) is given by the dotted red lines.

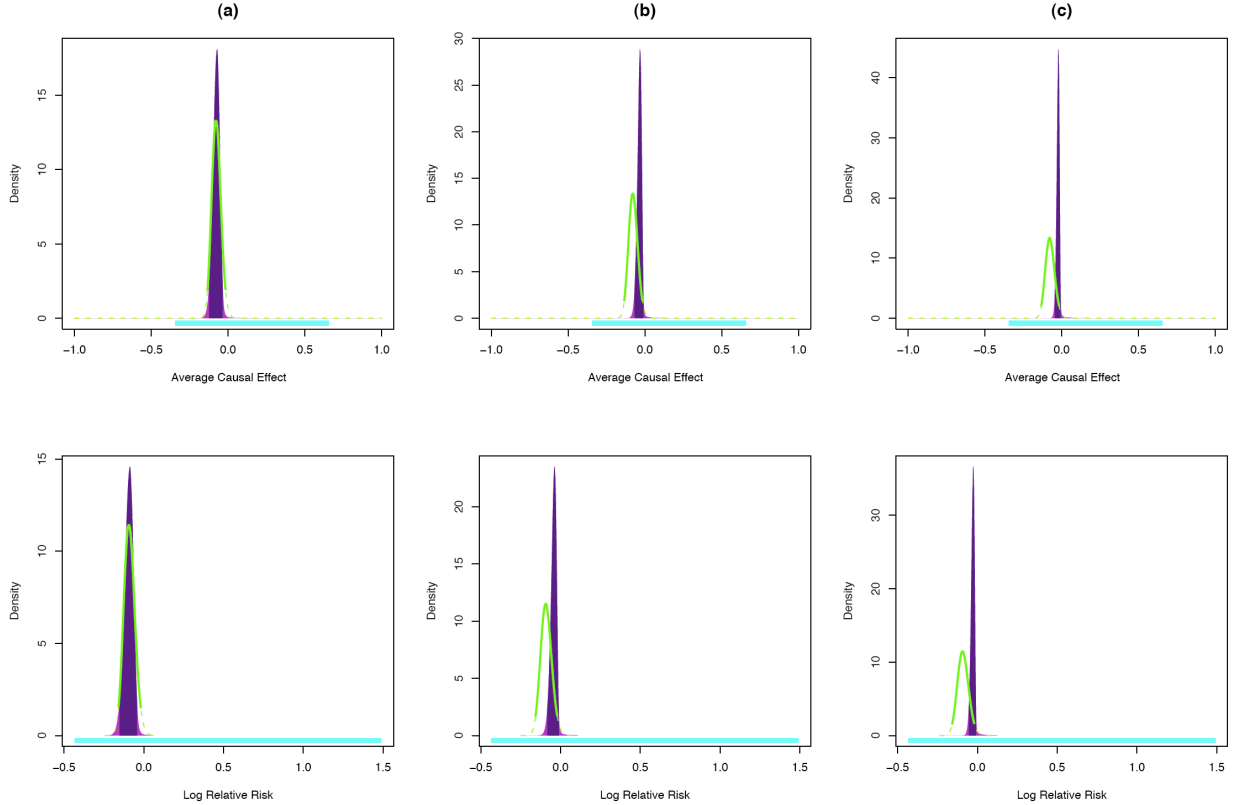


Figure 3: *Posterior Distributions of Prima Facie and Sensitivity Analysis Average Treatment Effects and Relative Risks for Data in Table 1 Under Informative Subjective Priors.* The green line in each panel is the posterior density of the prima facie ATE or RR (solid line depicts 95% highest posterior density region). The purple polygons depict the posterior density of the sensitivity analysis ATE or RR under alternative priors for ψ (dark purple denotes the 95% highest posterior density region). The horizontal light blue line in each each figure denotes the region inside the large sample nonparametric bounds on the estimand. The leftmost panels (column (a)) are from the model in which $b_{00} = 0.02, c_{00} = 10, b_{01} = 17, c_{01} = 5, b_{10} = 5, c_{10} = 17, b_{11} = 10,$ and $c_{11} = 0.02$. The center panels (column (b)) are from the model in which $b_{00} = 0.02, c_{00} = 10, b_{01} = 25, c_{01} = 3, b_{10} = 3, c_{10} = 25, b_{11} = 10,$ and $c_{11} = 0.02$. Finally, the rightmost panels (column (c)) are from the model that assumes $b_{00} = 0.02, c_{00} = 10, b_{01} = 45, c_{01} = 3.5, b_{10} = 3.5, c_{10} = 45, b_{11} = 10,$ and $c_{11} = 0.02$

In this application it seems relatively uncontroversial to assume that perceiving jury lists to be drawn from voter lists should not cause anyone to vote who would not have voted absent such a belief. In other words, it seems safe to assume a negative monotonic treatment effect. Put still another way, we do not anticipate a *jury attraction* effect. At most one would expect that only a minuscule fraction of the population is helped by the treatment variable. We operationalize this belief by assuming $\psi_{00} \sim \text{Beta}(0.02, 10)$ and $\psi_{11} \sim \text{Beta}(10, 0.02)$. Under such a prior, the

probability that the percentage of helped individuals in the $(X = 0, Y = 0)$ group is less than 2% is 0.975. The same is true for the percentage of helped individuals in the $(X = 1, Y = 1)$ group. As we will see below, this relatively uncontroversial assumption allows us to greatly narrow down the range of plausible ATE and RR values.

It seems that a wider range of defensible subjective beliefs are possible about ψ_{01} and ψ_{10} . Recall that ψ_{01} is the fraction of always succeed individuals and $(1 - \psi_{01})$ is the fraction of hurt individuals in the $(X = 0, Y = 1)$ group, while ψ_{10} is the fraction of hurt individuals and $(1 - \psi_{10})$ is the fraction of never succeed individuals in the $(X = 1, Y = 0)$ group. Researchers who hypothesize that there is a large jury aversion effect will believe that ψ_{01} is relatively small and ψ_{10} is relatively large. On the other hand, researchers who think that such an effect is small or essentially nonexistent will believe that ψ_{01} is close to 1 and ψ_{10} is close to 0. We operationalize these beliefs, along with beliefs in between these extremes with three prior distributions. We label these priors (a), (b), and (c). Prior (a) represents a belief that there is a fairly substantial jury aversion effect, (b) is a belief in a moderate effect, and (c) is a belief in a fairly small effect. Figure 2 plots the prior densities for each element of ψ under each set of prior beliefs. Note that in all three priors we assume an essentially negative monotonic treatment effect by assuming $\psi_{00} \sim \text{Beta}(0.02, 10)$ and $\psi_{11} \sim \text{Beta}(10, 0.02)$.

Under prior (a) we assume $\psi_{01} \sim \text{Beta}(17, 5)$ and $\psi_{10} \sim \text{Beta}(5, 17)$. This is consistent with a belief that there is a 95% chance that the fraction of hurt individuals in the $(X = 0, Y = 1)$ group is between 0.082 and 0.419. The same is true for the fraction of hurt individuals in the $(X = 1, Y = 0)$ group. In each case, the median fraction of hurt individuals is believed to be 0.219. We think very few researchers believe that more than about 40% of the individuals in these groups would be (or were) swayed to abstain from registering to vote solely by a belief that registering may make them slightly more likely to serve jury duty. Similarly, the lower bound of about 8% is consistent with a small but noticeable jury aversion effect. We think prior (a) is a reasonable representation of the beliefs of someone who believes in a fairly sizable, but still plausible jury aversion effect.

Prior (c) is designed to represent the beliefs of someone much more skeptical of any such effect.

Here we assume that $\psi_{01} \sim \text{Beta}(45, 3.5)$ and $\psi_{10} \sim \text{Beta}(3.5, 45)$. This is consistent with a belief that there is a 95% chance that the fraction of hurt individuals in the $(X = 0, Y = 1)$ group is between 0.018 and 0.159. The same is true for the fraction of hurt individuals in the $(X = 1, Y = 0)$ group. The prior median of each of these fractions is 0.066. While this prior assumes that the fractions of individual who are hurt by treatment is probably less than about 15% of each of the two relevant treatment-outcome groups it is still consistent with a negative jury aversion effect. Indeed, the expected fractions of hurt individuals in each of the two relevant outcome-treatment groups is about 7% which would not seem to be completely negligible.

Prior (b) falls somewhere between priors (a) and (c). Here we assume that $\psi_{01} \sim \text{Beta}(25, 3)$ and $\psi_{10} \sim \text{Beta}(3, 25)$. This is consistent with a belief that there is a 95% chance that the fraction of hurt individuals in the $(X = 0, Y = 1)$ group is between 0.024 and 0.243. The same is true for the fraction of hurt individuals in the $(X = 1, Y = 0)$ group. The prior median of each of these fractions is 0.098.

Results from analyses under priors (a), (b), and (c) are plotted in Figure 3. The first thing that is apparent from this figure is that the subjective priors have greatly decreased the plausible range of the causal effects of interest. By assuming an essentially negative monotonic treatment effect we forced most of the posterior mass to be to the left of 0. Given that the large sample nonparametric bounds were asymmetric around 0 with more room to the right of 0, this prior assumption adds a lot of information in a way that we think is relatively uncontroversial.

Where exactly to the left of 0 most of the posterior mass lies depends on one's prior beliefs about ψ_{01} and ψ_{10} . Under prior (a) we see that the posterior mean of ATE_s is -0.077 with 95% HPD region $[-0.124, -0.035]$. This is quite similar to the prima facie ATE. Under prior (b) the posterior mean of ATE_s is -0.035 with 95% HPD region $[-0.067, -0.010]$. Finally, under prior (c) we have that the posterior mean of ATE_s is -0.023 with 95% HPD region $[-0.046, -0.006]$. From this analysis we would conclude that the prima facie results derived from the simple 2×2 table can be supported with not implausible beliefs about unmeasured confounding. However, such

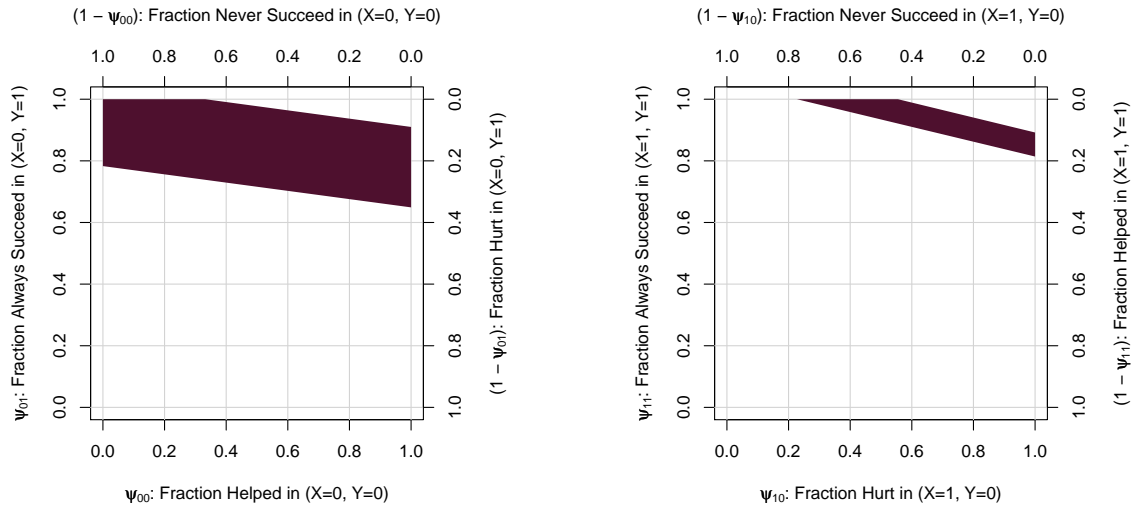


Figure 4: Values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} for which $|\Pr_p(Y(X = x) = y) - \Pr_s(Y(X = x) = y)| \leq 0.025$ for $x = 0, 1$ and $y = 0, 1$ when $\theta_{00} = 0.026, \theta_{01} = 0.191, \theta_{10} = 0.152, \theta_{11} = 0.631$. These θ values correspond to the posterior mean values of these parameters given the data in Table 1 and a non-informative prior for θ . The shaded regions depict values that satisfy the inequality above.

beliefs are at the extreme of what we think are reasonable. Other reasonable beliefs about potential unmeasured confounding yield much smaller (and substantively unimportant) jury aversion effects.

6.3 Ranges of ψ Consistent with the Estimated Prima Facie Post-Intervention Distribution

A more comprehensive picture of what beliefs about ψ can support inferences close to the prima facie inferences can be obtained by looking at the confounding plot discussed in Section 5. Figure 4 plots the values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} for which

$|\Pr_p(Y(X = x) = y) - \Pr_s(Y(X = x) = y)| \leq 0.025$ for $x = 0, 1$ and $y = 0, 1$ when θ is set to its posterior mean value.

Here we see that inferences are relatively insensitive to beliefs about ψ_{00} —any value of ψ_{00} between 0 and 1 can produce a sensitivity analysis post-intervention distribution that is close to the prima facie post-intervention distribution. This should not be unexpected given the relatively small fraction of observations that fall in the $(X = 0, Y = 0)$ cell. However, the prima facie

inferences are quite sensitive to beliefs about ψ_{11} — the fraction of always succeed individuals in the $(X = 1, Y = 1)$ cell. This parameter must be between about 0.8 and 1 in order for the sensitivity analysis post intervention distribution to be close to its prima facie counterpart. As we argued above, it seems quite reasonable to assume that ψ_{11} is very close to 1 in this application (there is essentially no jury attraction effect) so this sensitivity to ψ_{11} may not be so important. More important is the sensitivity of the prima facie results to the values of ψ_{01} and ψ_{10} . Here we see that the fraction of hurt individuals in the $(X = 0, Y = 1)$ cell cannot be too great while the fraction of hurt individuals in the $(X = 1, Y = 0)$ cell has to be moderately large.

6.4 Dealing with a Measured Confounder: Bayesian Analysis of a $2 \times 2 \times 5$ Table

We can also use Bayesian methods to model general unmeasured confounding while simultaneously adjusting for measured confounders. Here we deal with the situation in which there is a single measured confounder. However, the same techniques could, in principle, be used to deal with multiple measured confounders that are all discrete.

Table 4 presents the data in Table 1 broken down by occupational category ($W = 0$: self-employed, $W = 1$: hourly workers, $W = 2$: others working, $W = 3$: not in workforce, $W = 4$: retired). Since the costs associated with serving jury duty are, to a large extent, tied to one's occupation and since occupational status is believed to have some effect on political participation we expect occupation to act as a confounder and it would seem sensible to adjust for this measured confounder. Such an adjustment would be especially useful if we thought it removed most of the confounding bias (unlikely in this application) and/or if it made it easier to reason about the unobserved potential outcomes (somewhat true in this case). In situations where adjustment for the measured confounding variables is thought to remove only a small portion of the confounding bias it is not always clear whether the additional information provided by conditioning on the measured confounding variables outweighs the additional complication of formulating a prior conditional on X , Y , and the measured confounders.

If we are willing to assume the measured confounder W is discrete with K categories it is appropriate to perform the simple 2×2 analyses described above separately within each of the $k = 0, \dots, K - 1$ levels of W and then weight by $\phi_k \equiv \Pr(W = k)$. Somewhat more formally, we can write the likelihood for the observed data marginalized over the unobserved Z variables as:

$$\begin{aligned}
p(\mathbf{w}, \mathbf{x}, \mathbf{y} | \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^n \sum_{z_i \in \mathcal{Z}_i} p(w_i, x_i, y_i, z_i | \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) \\
&= \prod_{i=1}^n p(w_i, x_i, y_i | \boldsymbol{\phi}, \boldsymbol{\theta}) \left\{ \sum_{z_i \in \mathcal{Z}_i} p(z_i | w_i, x_i, y_i, \boldsymbol{\psi}) \right\} \\
&= \prod_{i=1}^n p(w_i | \boldsymbol{\phi}) p(x_i, y_i | w_i, \boldsymbol{\theta}) \left\{ \sum_{z_i \in \mathcal{Z}_i} p(z_i | w_i, x_i, y_i, \boldsymbol{\psi}) \right\}
\end{aligned}$$

where the $p(w_i | \boldsymbol{\phi})$ term is a multinomial mass function with probability vector $\boldsymbol{\phi}$ and multinomial sample size 1, $p(x_i, y_i | w_i, \boldsymbol{\theta})$ is a multinomial mass function that depends on the value of w_i with probabilities $\theta_{00|w_i}, \theta_{01|w_i}, \theta_{10|w_i}, \theta_{11|w_i}$ and multinomial sample size 1, and the term involving the sum over Z consists of $4K$ Bernoulli mass functions with parameters $\psi_{00|w_i}, \psi_{01|w_i}, \psi_{10|w_i}$ and $\psi_{11|w_i}$. Note that everything is analogous to the 2×2 analysis except for the conditioning on w_i throughout.

Now that we are explicitly adjusting for W the prima facie post-intervention distribution must

be redefined to account for this adjustment. It becomes:

$$\begin{aligned}
\Pr_{pw}(Y(X = 0) = 0) &= \sum_{w=0}^{K-1} \Pr(Y = 0|X = 0, W = w) \Pr(W = w) \\
&= \sum_{w=0}^{K-1} \frac{\theta_{00|w}}{\theta_{00|w} + \theta_{01|w}} \phi_w \\
\Pr_{pw}(Y(X = 0) = 1) &= \sum_{w=0}^{K-1} \Pr(Y = 1|X = 0, W = w) \Pr(W = w) \\
&= \sum_{w=0}^{K-1} \frac{\theta_{01|w}}{\theta_{00|w} + \theta_{01|w}} \phi_w \\
\Pr_{pw}(Y(X = 1) = 0) &= \sum_{w=0}^{K-1} \Pr(Y = 0|X = 1, W = w) \Pr(W = w) \\
&= \sum_{w=0}^{K-1} \frac{\theta_{10|w}}{\theta_{10|w} + \theta_{11|w}} \phi_w \\
\Pr_{pw}(Y(X = 1) = 1) &= \sum_{w=0}^{K-1} \Pr(Y = 1|X = 1, W = w) \Pr(W = w) \\
&= \sum_{w=0}^{K-1} \frac{\theta_{11|w}}{\theta_{10|w} + \theta_{11|w}} \phi_w
\end{aligned}$$

Similarly, the sensitivity analysis post-intervention distribution becomes:

$$\begin{aligned}
\Pr_{sw}(Y(X=0)=0) &= \sum_{w=0}^{K-1} \sum_{z=0}^3 \Pr(Y=0|X=0, W=w, Z=z) \Pr(Z=z|W=w) \Pr(W=w) \\
&= \sum_{w=0}^{K-1} [\theta_{10|w}(1-\psi_{10|w}) + \theta_{11|w}(1-\psi_{11|w}) + \theta_{00|w}] \phi_w \\
\Pr_{sw}(Y(X=0)=1) &= \sum_{w=0}^{K-1} \sum_{z=0}^3 \Pr(Y=1|X=0, W=w, Z=z) \Pr(Z=z|W=w) \Pr(W=w) \\
&= \sum_{w=0}^{K-1} [\theta_{10|w}\psi_{10|w} + \theta_{11|w}\psi_{11|w} + \theta_{01|w}] \phi_w \\
\Pr_{sw}(Y(X=1)=0) &= \sum_{w=0}^{K-1} \sum_{z=0}^3 \Pr(Y=0|X=1, W=w, Z=z) \Pr(Z=z|W=w) \Pr(W=w) \\
&= \sum_{w=0}^{K-1} [\theta_{00|w}(1-\psi_{00|w}) + \theta_{01|w}(1-\psi_{01|w}) + \theta_{10|w}] \phi_w \\
\Pr_{sw}(Y(X=1)=1) &= \sum_{w=0}^{K-1} \sum_{z=0}^3 \Pr(Y=1|X=1, W=w, Z=z) \Pr(Z=z|W=w) \Pr(W=w) \\
&= \sum_{w=0}^{K-1} [\theta_{00|w}\psi_{00|w} + \theta_{01|w}\psi_{01|w} + \theta_{11|w}] \phi_w
\end{aligned}$$

Prima facie and sensitivity analysis average treatment effects and relative risks are defined in the usual way from these post-intervention distributions. After a bit of algebra it can be shown that the large sample nonparametric bounds for the *ATE* and *RR* do not change when moving from the simple 2×2 analysis to the analysis in which a nonparametric adjustment is made for W and no additional conditional independence assumptions are made.

We perform two analyses that explicitly adjust for reported occupation. In the first analysis we assume a uniform prior for each of the 20 $\psi_{xy|w}$ parameters. The second analysis assumes $b_{00|w} = 0.02, c_{00|w} = 10, b_{01|w} = 25, c_{01|w} = 3, b_{10|w} = 3, c_{10|w} = 25, b_{11|w} = 10, c_{11|w} = 0.02$ for $w = 0, 1, 2$ and $b_{00|w} = 0.02, c_{00|w} = 10, b_{01|w} = 15, c_{01|w} = 1, b_{10|w} = 1, c_{10|w} = 15, b_{11|w} = 10, c_{11|w} = 0.02$ for $w = 3, 4$. This informative subjective prior assumes that there is an essentially negative monotonic treatment effect and that the largest effects are found among those who are working ($w = 0, 1, 2$). In both analyses we assume non-informative priors for θ and ϕ . Figure 5 displays the results of

these analyses graphically.

Looking at these results we see that the posterior mean of ATE_{pw} is -0.082 with 95% HPD interval $[-.137 - 0.024]$ and the posterior mean of $\log(RR)_{pw}$ is -0.100 with 95% HPD interval $[-0.162, -0.028]$ note that these results are quite close to those for ATE_p and RR_p suggesting that, by itself, reported occupational category is not a serious confounder. Although it is worth noting that the effects adjusted for occupation are slightly larger in magnitude and the intervals have narrowed somewhat.

Under the uniform prior for each $\psi_{xy|w}$ we see that the posterior mean of ATE_{sw} is 0.156 with 95% HPD region $[-0.039, 0.351]$ and the posterior mean of $\log(RR)_{sw}$ is 0.250 with 95% HPD region $[-0.064, 0.586]$. Note the dramatic narrowing of the HPD intervals here relative to the intervals for from the uniform prior for ψ_{xy} unconditional on occupation. As we might expect, conditioning on a measured confounder tends to decrease the variability of one's estimates.

Looking at the results under the subjective prior on ψ we see that the posterior mean of ATE_{sw} is -0.031 with 95% HPD region $[-0.048, -0.016]$ and the posterior mean of $\log(RR)_{sw}$ is -0.038 with 95% HPD interval $[-0.058, -0.021]$. Again, these results are roughly consistent with those from the appropriate 2×2 analysis except that the sampling variability has decreased after adjusting for occupation.

It is reassuring to note that the results presented here based on justifiable subjective prior distributions correspond closely at a qualitative level to the results of Oliver and Wolfinger (1999) who adjusted for a number measured covariates that were available in the original dataset. While Oliver and Wolfinger do not explicitly calculate the same causal effects as we have in this paper they reach a similar qualitative conclusion. To cite from their conclusion:

There is a grain—but no more than a grain—of truth to the belief that jury aversion reduces voter registration... People who believe that registering makes them susceptible to jury service are slightly less likely to be registered than those who identify other sources. This aversion to jury service is responsible for a drop in voter turnout of less than one percentage point. (p. 151)

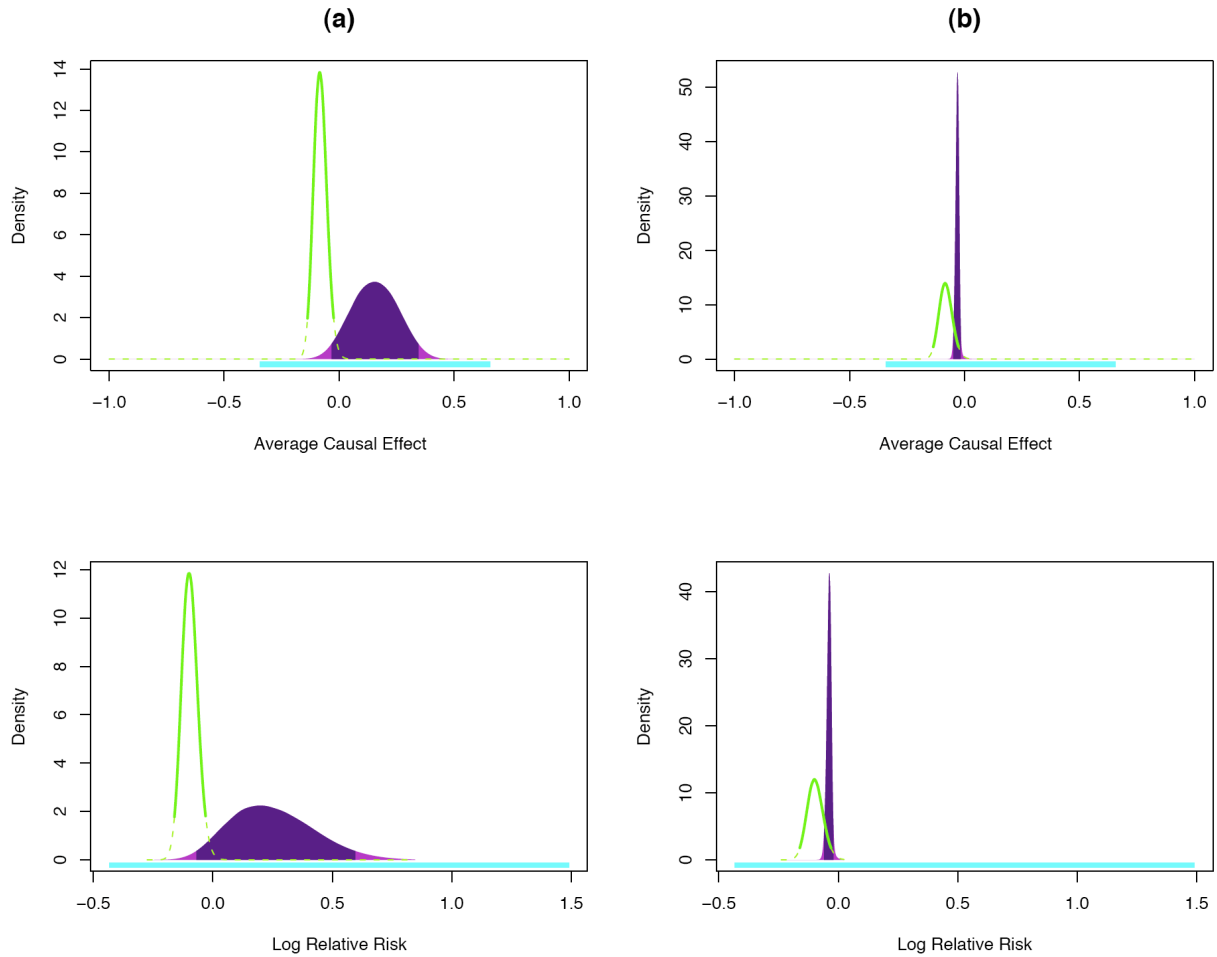


Figure 5: *Posterior Distributions of Prima Facie and Sensitivity Analysis Average Treatment Effects and Relative Risks for Data in Table 4.* The green line in each panel is the posterior density of the prima facie ATE or RR (solid line depicts 95% highest posterior density region). The purple polygons depict the posterior density of the sensitivity analysis ATE or RR under alternative priors for ψ (dark purple denotes the 95% highest posterior density region). The horizontal light blue line in each each figure denotes the region inside the large sample nonparametric bounds on the estimand. The leftmost panels (column (a)) are from a model in which a uniform prior is assumed for each element of ψ within each occupational category. The rightmost panels (column (b)) are from the model that assumes $b_{00|w} = 0.02, c_{00|w} = 10, b_{01|w} = 25, c_{01|w} = 3, b_{10|w} = 3, c_{10|w} = 25, b_{11|w} = 10, c_{11|w} = 0.02$ for $w = 0, 1, 2$ and $b_{00|w} = 0.02, c_{00|w} = 10, b_{01|w} = 15, c_{01|w} = 1, b_{10|w} = 1, c_{10|w} = 15, b_{11|w} = 10, c_{11|w} = 0.02$ for $w = 3, 4$. This informative subjective prior assumes that there is an essentially negative monotonic treatment effect and that the largest effects are found among those who are working ($w = 0, 1, 2$).

7 Discussion

All counterfactual causal inferences require fundamentally untestable causal assumptions. In a well-run randomized controlled experiment the necessary assumptions for point identification of causal effects are so plausible that they hardly appear to be assumptions. However, as one moves away from the experimental ideal the necessary ignorability assumptions typically become less plausible. Better research designs and improved estimation strategies can do much to help decrease the impact of confounding bias. Nonetheless, researchers still need to make causal assumptions in order to make causal inferences. The development of methods of sensitivity analyses for situations in which unmeasured confounding is present, as is done in this paper, serves to shift empirical social science research away from the all too typical enterprise of defending indefensible causal assumptions to the practice of honestly stating the range of assumptions that are consistent with a particular type of causal effect. *This has the potential to accelerate the accumulation of knowledge in the social sciences.*

While the running example in this paper dealt with a large sample survey of U.S. citizens, the methods discussed here are just as, if not more, relevant for qualitative researchers in comparative politics and international relations who typically deal with much smaller numbers of cases. Indeed, as Sekhon (2004) has pointed out, causal claims in such settings typically involve the same sorts of counterfactuals considered here (see also Fearon (1991) and Hawthorn (1993)). The methods presented here merely require researchers to specify their beliefs about four relevant counterfactual quantities— the fractions of “helped”, “hurt”, “always succeed”, and “never succeed” units within various observed combinations of X and Y . To the extent that the consideration of various counterfactuals is what researchers in these fields already do, the methods presented in this paper can be seen as a way to ensure that the conclusions reached by these scholars are logically consistent with the counterfactual claims they are willing to maintain. Further, because researchers in these fields are often unable to collect additional data on many important confounding variables they

often find themselves dealing with simple 2×2 and $2 \times 2 \times K$ contingency tables where substantial unmeasured confounding is thought to be present.

References

- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.
- Balke, Alexander, and Judea Pearl. 1997. "Bounds on Treatment Effects From Studies With Imperfect Compliance." *Journal of the American Statistical Association* 92(439):1171–1176.
- Bernardo, José, and Adrian F. M. Smith. 1994. *Bayesian Theory*. New York: Wiley.
- Brumback, Babette A., Miguel A. Hernán, Sebastien J. P. A. Haneuse, and James M. Robins. 2004. "Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures." *Statistics in Medicine* 23:749–767.
- Chickering, David Maxwell, and Judea Pearl. 1997. "A Clinician's Tool for Analyzing Non-Compliance." *Computing Science and Statistics* 29(2):424–431.
- Cornfield, J., W. Haenszel, E. Hammond, A. Lilienfeld, M. Shimkin, and E. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22:173–203.
- Fearon, James. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43:474–484.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. London: Chapman & Hall, second edition.
- Gill, Jeff. 2007. *Bayesian Methods: A Social and Behavioral Science Approach*. Boca Raton, FL: Chapman & Hall/CRC, second edition.
- Gill, Jeff, and Lee D. Walker. 2005. "Elicited Priors for Bayesian Model Specifications in Political Science Research." *The Journal of Politics* 67(3):841–872.
- Hawthorn, Geoffrey. 1993. *Plausible Worlds: Possibility and Understanding in History and the Social Sciences*. New York: Cambridge University Press.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- Imai, Kosuke, and Samir Soneji. 2007. "On the Estimation of Disability-Free Life Expectancy: Sullivan's Method and Its Extension." *Journal of the American Statistical Association* 102:1199–1211.
- Imai, Kosuke, and Teppei Yamamoto. 2008. "Causal Inference with Measurement Error: Nonparametric Identification and Sensitivity Analyses of a Field Experiment on Democratic Deliberations." Paper presented at the 2008 Summer Political Methodology Meeting.
- Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: an Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44(2):375–404.
- Kadane, Joseph B., and Lara J. Wolfson. 1998. "Experiences in Elicitation." *The Statistician* 47:3–19.
- King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347–361.
- King, Gary, and Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409–1427.
- Lin, D. Y., B. M. Psaty, and R. A. Kronmal. 1998. "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies." *Biometrics* 54:948–963.
- Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review Papers and Proceedings* 80(2):319–323.
- Manski, Charles F. 1993. "Identification Problems in the Social Sciences." *Sociological Methodology* 23:1–56.
- Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.
- Miller, Warren E., Donald R. Kinder, Steven J. Rosenstone, and the American National Election Studies. 1991. "American National Election Study: 1990-1991 Panel Study of the Political Consequences of War / 1991 Pilot Study." [computer file] (Study #9673) Conducted by the Center for Political Studies of the Institute for Social Research, the University of Michigan.
- Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.
- Oliver, J. Eric, and Raymond E. Wolfinger. 1999. "Jury Aversion and Voter Registration." *American Political Science Review* 93(1):147–152.
- Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669–710.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Richardson, Thomas. 2004. "Discussion on the Paper by Wakefield." *Journal of the Royal Statistical Society: Series A* 167:438.
- Robins, James M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.
- Robins, James M. 1989. "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies." In *Health Service Research Methodology: A Focus on AIDS* (L Sechrest, H. Freeman, and A. Mulley, editors), Washington DC: U.S. Public Health Service.
- Robins, James M. 1999. "Association, Causation, and Marginal Structural Models." *Synthese* 121:151–179.
- Robins, James M. 2002. "Comment." *Statistical Science* 17:309–321.
- Rosenbaum, Paul R. 1987. "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies." *Biometrika* 74:13–26.

- Rosenbaum, Paul R. 2002a. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Statistical Science* 17:286–304.
- Rosenbaum, Paul R. 2002b. *Observational Studies*. New York: Springer, second edition.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983a. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society: Series B* 45:212–218.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.
- Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment." *Journal of the American Statistical Association* 75(371):591–593.
- Rubin, Donald B. 1986. "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81(396):961–962.
- Schlesselman, James J. 1978. "Assessing Effects of Confounding Variables." *American Journal of Epidemiology* 108:3–8.
- Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals." *Perspectives on Politics* 2(2):281–293.
- Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(June):412–423.

$W = 0$ Self-Employed

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	1	21
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	10	93

$W = 1$ Hourly Workers

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	5	32
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	27	92

$W = 2$ Others Working

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	4	44
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	52	186

$W = 3$ Not in Workforce

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	7	20
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	19	47

$W = 4$ Retired

	$Y = 0$ Did <i>Not</i> Register to Vote	$Y = 1$ Did Register to Vote
$X=0$ Perceived Source of Jury Lists <i>Does Not</i> Include Voter Lists	2	26
$X=1$ Perceived Source of Jury Lists Does Include Voter Lists	6	55

Table 4: *Perceived Source of Jury Lists and Voter Registration Among Citizens With Some Self Reported Knowledge of How Jury Lists are Constructed by Occupational Category.* Each entry is the number of citizens in that category. Data from Table 2 of Oliver and Wolfinger (1999).