Statistical Science
2017, Vol. 32, No. 4, 561–579
DOI: 10.1214/17-STS623

Instrumental Variable Estimation with a Stochastic Monotonicity Assumption

Dylan S. Small, Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch and M. Alan Brookhart

Abstract. The instrumental variables (IV) method provides a way to estimate the causal effect of a treatment when there are unmeasured confounding variables. The method requires a valid IV, a variable that is independent of the unmeasured confounding variables and is associated with the treatment but which has no effect on the outcome beyond its effect on the treatment. An additional assumption often made is deterministic monotonicity, which says that for each subject, the level of the treatment that a subject would take is a monotonic increasing function of the level of the IV. However, deterministic monotonicity is sometimes not realistic. We introduce a stochastic monotonicity assumption, a relaxation that only requires a monotonic increasing relationship to hold across subjects between the IV and the treatments conditionally on a set of (possibly unmeasured) covariates. We show that under stochastic monotonicity, the IV method identifies a weighted average of treatment effects with greater weight on subgroups of subjects on whom the IV has a stronger effect. We provide bounds on the global average treatment effect under stochastic monotonicity and a sensitivity analysis for violations of stochastic monotonicity. We apply the methods to a study of the effect of premature babies being delivered in a high technology neonatal intensive care unit (NICU) vs. a low technology unit.

Key words and phrases: Causal inference, observational study, instrumental variable, two stage least squares.

1. INTRODUCTION

The instrumental variable (IV) method provides a way to learn about the effect of a treatment when there is unmeasured confounding under certain assumptions. There are several approaches to using IVs and corresponding definitions for an IV (e.g., Angrist, Imbens and Rubin, 1996, Hernán and Robins, 2006), however, all the definitions include that an IV satisfies the core assumptions of (1) the IV is associated with the treatment; (2) the IV is independent of unmeasured confounders; and (3) the IV only affects the outcome through its effect on treatment received (the exclusion restriction). In addition, two basic causal assumptions that are commonly made are (i) there is no interference between units; and (ii) there are no unrepresented versions of the treatment, which means there are not different ways of administering the same level of the treatment that lead to different outcomes for a given subject (Rubin, 1986). These core IV assumptions and

Dylan S. Small is Class of 1965 Wharton Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: dsmall@wharton.upenn.edu). Zhiqiang Tan is Professor, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA (e-mail: ztan@stat.rutgers.edu). Roland R. Ramsahai is Vice President at Reinsurance Division of Berkshire Hathaway Group, Stamford, CT 06902 (e-mail: roland.ramsahai@hotmail.com). Scott A. Lorch is Associate Professor, Department of Pediatrics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: LORCH@email.chop.edu). M. Alan Brookhart is Professor, Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina 27599, USA (e-mail: mabrook@email.unc.edu).

basic causal assumptions provide bounds for the causal effects of treatment, but additional assumption(s) are needed to point identify a causal effect (Robins, 1989, Manski, 1990, Balke and Pearl, 1997, Angrist, Imbens and Rubin, 1996). Angrist, Imbens and Rubin (1996) introduced two additional assumptions: (a) no unrepresented versions of the IV—there are not different ways of administering the same level of the IV that lead to different treatment for a given subject and (b) deterministic monotonicity—for a binary treatment and a binary IV, there are no defier subjects who would take the control if encouraged to take the treatment by the IV but take the treatment if not encouraged by the IV (the name deterministic monotonicity comes from that for all subjects, the level of the treatment that a subject would take is a monotone increasing function of the IV). Under these additional assumptions as well as the basic causal and core IV assumptions, Angrist, Imbens and Rubin (1996) showed that the Wald estimate (the difference in the mean outcome between subjects with the encouraging level of the IV vs. the not encouraging level divided by the difference in the proportion of subjects taking the treatment with the encouraging level of the IV vs. not the encouraging level) converges to the average treatment effect for compliers, where the compliers are the subjects who would take the treatment if encouraged by the IV but not take the treatment if not encouraged by the IV; the average treatment effect for compliers is also called the local average treatment effect (LATE).

The additional assumptions of (a) no unrepresented versions of the IV and (b) deterministic monotonicity are plausible for a setting where the IV is delivered in a uniform way and the encouraging level of the IV provides a clear incentive to take the treatment with no disincentive (Imbens, 2014, Section 5.3). For example, consider Finkelstein et al.'s (2012) study of the effect of having the effect of having health insurance on future health. Finkelstein et al. (2012) considered low-income adults in Oregon, where in 2008 a waiting list was opened for a state Medicaid expansion program, which had been closed to new enrollment since 2004. Because the waiting list exceeded the number of spots available, the state drew names by lottery to decide who would have the opportunity to enroll in Medicaid. Not all winners enrolled in Medicaid either because they did not apply or were deemed ineligible, and some non-winners obtained health insurance through other means; however, winners had about a 25 percentage point higher chance of having health insurance one year after the lottery. Finkelstein et al. (2012) used winning the lottery vs. not winning as an IV and inferred that health insurance improved self-reported physical and mental health. The no unrepresented versions of the IV assumption is plausible because the IV is delivered in a uniform way—if a person receives the encouraging level (winning the lottery), she is provided the opportunity to obtain health insurance by enrolling in Medicaid while if a person receives the non-encouraging level (not winning the lottery), she is denied the opportunity to enroll in Medicaid. The deterministic monotonicity assumption is plausible because winning the lottery provides a clear incentive to obtain health insurance (free enrollment in Medicaid) with no disincentive.

Although the additional assumptions of (a) no unrepresented versions of the IV and (b) deterministic monotonicity are plausible for some studies using an IV, there are many studies for which they are not plausible, particularly when the high level of the IV provides incentives to take the treatment but at the same time provides certain disincentives. For example, consider a study of the effect of mortality for premature babies of being delivered in a high level neonatal intensive care unit (high level NICU) vs. a low level NICU—a high level NICU is a NICU that has the capacity for sustained mechanical assisted ventilation and delivers at least 50 premature babies per year (Lorch et al., 2012a). Yang, Lorch and Small (2014) used an IV approach where the IV was whether or not the mother's excess travel time from the nearest high level NICU compared to the nearest low level NICU (under average traffic conditions) is less than or equal to 10 minutes. Excess travel time to a specialty care facility compared to a normal facility has been used as an IV for whether a person receives a certain type of care in many health studies, for example, McClellan, McNeil and Newhouse (1994). In the NICU study, living near to a high level NICU strongly encourages a mother to deliver at a high level NICU: in data from Pennsylvania (see Section 8), 75% of mothers who live relatively near to a high level NICU (excess travel time $\leq$10 minutes) deliver at a high level NICU whereas only 30% of mothers who live relatively far from a high level NICU (excess travel time $>$10 minutes) deliver at a high level NICU. However, the excess travel time IV is not delivered in a uniform way. Excess travel time is a function of where a mother lives, which may influence the choice of hospital in ways other than through excess travel time, such as community, family and friends' views of hospitals in the area. Different places that are both relatively near to a high level NICU (excess travel

TABLE 1
Comparison of four zip codes' demographics, IV (excess travel time) values and proportion of deliveries at a high level NICU

Zip Code	*A*	*B*	*C*	*D*
Proportion High School Degree	0.80	0.77	0.86	0.84
Proportion College Degree	0.16	0.10	0.18	0.15
Average Income	$33,914	$34,534	$34,802	$32,896
Rural Urban Continuum Code*	2	2	4	4
IV $Z^{\dagger}$	1	0	1	0
Proportion of deliveries at high level NICU	0.28	0.74	0.94	0.14

*Code of 2 means zip code is in county that is part of metropolitan area of 200,000 to 1 million people; 4 means zip code is in county that is adjacent to but not part of a metropolitan area and county has an urban population of at least 20,000.
$^{\dagger}Z = 1(0)$ means excess travel time of ≤ 10 (>10) minutes.

time ≤ 10 minutes) are unrepresented versions of the IV and there might be defiers with respect to some versions of the IV. For example, consider zip codes *A* and *B* shown in Table 1. The two zip codes are demographically similar. Zip code *A* (a zip code in the Wilkes Barre, PA area) is relatively near to a high level NICU with an excess travel time of only 6 minutes while zip code *B* (a zip code in the Erie, PA area) is relatively far from a high level NICU with an excess travel time of 30 minutes. With respect to the IV of ≤ 10 vs. >10 minutes excess travel time, if there are no defiers for the versions of the IV given by zip codes *A* and *B*, then the proportion of mothers delivering at a high level NICU should be at least as high in *A* as in *B*. But in fact, only 28% of mothers in the near zip code *A* deliver at a high level NICU whereas 74% of mothers in the far zip code *B* deliver at a high level NICU (*p*-value from Fisher's exact test for high level NICU rate being lower in *A* than *B* is $<10^{-15}$). Thus, with respect to the versions of the IV represented by these two zip codes, there appear to be defiers. However, a mother who was a defier with respect to these two zip codes, that is, a mother who would deliver at a low level NICU if she lived in the near zip code *A* but would deliver at a high level NICU if she lived in the far zip code *B*, might be a complier with respect to two other zip codes. For example, consider zip codes *C* (in Dubois, PA area) and *D* (in Sugarcreek, PA) in Table 1. These two zip codes are demographically similar, but *C* is a near zip code and *D* is a far zip code. Almost all mothers in *C* deliver at a high level NICU whereas almost all mothers in *D* deliver at a low level NICU. Thus, almost all mothers are compliers with respect to zip code *C* vs. zip code *D*. In summary, there are unrepresented versions of the IV and violations of deterministic monotonicity for the NICU study.

1.1 Stochastic Monotonicity Assumption and Main Results

We consider a weakening of deterministic monotonicity called stochastic monotonicity. Informally, the stochastic monotonicity assumption says that if we stratify on all the measured and unmeasured confounders of the relationship between the treatment and outcome, then within each stratum, the probability of taking the treatment for subjects given the encouraging level of the IV is at least as high as for subjects given the non-encouraging level. Stochastic monotonicity is weaker than deterministic monotonicity because it does not require a monotonic relationship between IV and treatment for each subject but only a monotonic relationship between IV and probability of treatment within strata.

We show that if an IV satisfies stochastic monotonicity along with the basic causal and core IV assumptions, the IV can be used to learn about certain useful quantities even if deterministic monotonicity is violated. First, we show that the Wald estimate identifies a strength-of-IV weighted average treatment effect (SIV-WATE), a weighted average of treatment effects where conditionally on all possible confounders, the weight for a group of subjects is proportional to the size of the group times how much higher the probability of taking the treatment for subjects in the group when given the encouraging level of the IV is compared to the non-encouraging level (Proposition 1). This implies the no sign reversal property—if the sign of the treatment effects $(+, 0, -)$ is the same for all subjects, then this sign is identified by the IV (Section 4.3). Second, we show that the observable characteristics of the weighted population in the definition of the SIV-WATE are identified so the treatment effect

underlying the Wald estimate can be understood under stochastic monotonicity (Proposition 2). Third, we show that the stochastic monotonicity assumption can narrow bounds on the global average treatment effect, that is, the average treatment effect for the whole population (Proposition 3). These results are comparable to previous results that the deterministic monotonicity assumption can narrow the bounds on the global average treatment effect obtained under the core IV assumptions (Balke and Pearl, 1997). Fourth, we provide a method of sensitivity analysis for bias from violations of stochastic monotonicity that is analogous to Angrist, Imbens and Rubin's (1996) sensitivity analysis for bias from violations of deterministic monotonicity.

1.2 Review of Literature and Contributions of This Paper

Previous literature has considered relaxing some of the assumptions of the Angrist, Imbens and Rubin (1996) framework that are violated when the IV is not delivered in a uniform way or deterministic monotonicity is violated. Hernán and Robins (2006) and Chalak (2017) have presented results for when the measured IV is a proxy for an underlying, possibly continuous IV; their results assume that the underlying IV satisfies deterministic monotonicity whereas we allow for the underlying IV to violate deterministic monotonicity and only satisfy the weaker condition of stochastic monotonicity. When the IV is not a proxy and does not have unrepresented versions, several authors have presented results that identify the LATE or variants of the LATE when deterministic monotonicity is violated. DiNardo and Lee (2011) presented a stochastic monotonicity condition. Brookhart and Schneeweiss (2007) presented a similar formula when heterogeneity of treatment effects is generated by an unmeasured variable. Ramsahai (2012) presented bounds on the treatment effect for a binary outcome under a stochastic monotonicity condition. de Chaisemartin (2017) provided a relaxation of deterministic monotonicity under which if there is a subgroup of compliers that accounts for the same proportion as the defiers and that has the same average treatment effect as the defiers (called the "compliers-defiers" assumption), then the Wald estimate captures the average treatment effect of the remaining part of the compliers. Angrist, Imbens and Rubin (1996) presented a formula for the sensitivity of the Wald estimate to violations of deterministic monotonicity. Klein (2010) introduced local violations of deterministic monotonicity that are independent of unmeasured confounders and showed that the bias of the Wald estimate can be well approximated if such violations are small. Huber and Mellace (2012) considered a local monotonicity assumption which requires that there be only compliers or defiers conditional on each value of the outcome. Finally, Robins (1994), Hernán and Robins (2006), and Tan (2010) developed an IV approach that does not assume deterministic monotonicity but instead makes homogeneity or parametric heterogeneity assumptions about causal effects in different subgroups to achieve identification of average treatment effects on the treated at different instrument levels.

The contributions of this paper are that we provide a unified framework, identification results and inference methods that address simultaneously the problems that arise when an IV may not be delivered in a uniform way, the IV may be a proxy and the IV may violate deterministic monotonicity. Previous literature to our knowledge has not considered the presence of these three problems simultaneously. We show that the Wald estimator identifies a weighted average treatment effect, the SIV-WATE, even if deterministic monotonicity is violated so long as stochastic monotonicity is satisfied. In comparison to existing results on IV estimation without monotonicity, the stochastic monotonicity assumption we consider does not require there to be only compliers or defiers conditional on each value of the outcome as in Huber and Mellace (2012) or for the violations of deterministic monotonicity to be independent of unmeasured confounders as in Klein (2010). The stochastic monotonicity assumption we consider implies the "compliers-defiers" assumption of de Chaisemartin (2017); the advantage of the stochastic monotonicity assumption when it holds is that further insight into whom the Wald IV estimator pertains to is available, see for example Section 6 for discussion of different ways to interpret the SIV-WATE, the estimand of the Wald IV estimator under stochastic monotonicity.

Our paper is organized as follows. Section 2 provides notation and framework. Section 3 reviews the deterministic compliance class framework and identification results. Section 4 presents the stochastic compliance class framework, the stochastic monotonicity assumption and identification results. Section 5 provides sensitivity analysis for violations of stochastic monotonicity. Section 6 discusses interpretation of the treatment effect estimated by the Wald estimator under stochastic monotonicity. Section 7 discusses conditioning on covariates. Section 8 presents an application to the NICU study. Section 9 discusses stochas-

tic monotonicity for another example, physician prescribing preference IVs. Section 10 provides discussion. The proofs are provided in the supplementary materials (Small et al., 2017).

2. NOTATION AND FRAMEWORK

Let Z be the measured IV, D the treatment and Y the outcome. We will not consider covariates for the moment but discuss them in Section 7. The IV Z and the treatment D are assumed to be binary. We extend the results to non-binary Z in the supplementary materials. We refer to level 1 of the IV Z as the encouraging level and 0 as the non-encouraging level, and we refer to level 1 of D as the treatment and 0 as the control. There are N units (subjects). The notation $A \perp\!\!\!\perp B$ means the random variables A and B are independent and $A \perp\!\!\!\perp B|C$ means A and B are conditionally independent given C.

We make the following commonly made basic causal assumptions:

BA1 *No interference*: The observation on one subject is not affected by the assignment of treatments to the other subjects (Cox, 1958, Rubin, 1986).

BA2 *No unrepresented versions of treatment*: The treatment levels 1 and 0 adequately represent all versions of the treatment (Rubin, 1986). In other words, if there are two different ways of receiving the treatment that are both represented by level 1 (0), the potential outcomes corresponding to these two different ways of receiving the treatment are the same.

We will now state "core" assumptions for Z to be a valid IV; we call these core assumptions because they follow from the commonly used informal description of an IV as a variable that affects the treatment (relevance), but only affects the outcome by altering the treatment (exclusion restriction) and is independent of unmeasured confounders (effective randomness of the IV) (Hernán and Robins, 2006, Rosenbaum, 2010). We will state these core assumptions for three different settings: (i) Z is a causal IV that has a causal effect on the treatment; (ii) Z is an intensity preserving proxy for a causal IV Z^* that has a causal effect on the treatment; (iii) Z or something that Z is a proxy for cannot easily be thought of as being manipulated. We will then show in Section 2.4 that the three sets of core assumptions imply a set of common implications for the relationship between Z and the treatment and potential outcomes. Our subsequent results will only depend on these common implications holding.

A reader interested in our main results about stochastic monotonicity without the background on the various scenarios in which they are plausible and how they relate to existing results could skip ahead to Section 2.4 and then Section 4.

2.1 Core Assumptions for Z Being a Causal IV

The following are the three core assumptions for Z to be a valid causal IV (Hernán and Robins, 2006):

CA1-1 *Positive Causal Effect of Z on Treatment*: Let $D_i(z)$ be the treatment that subject i would receive if she were assigned level z of Z. Each subject has a set of potential treatments, $\{D_i(1), D_i(0)\}$. The positive causal effect of Z on the treatment assumption is that $E[D(1)] > E[D(0)]$.

CA2-1 *Exclusion Restriction*: Let $Y_i(z, d)$ be the potential outcome that subject i would experience if the causal IV Z is set to level z and the treatment is set to level d. The exclusion restriction is that $Y_i(0, d) = Y_i(1, d)$ for $d = 0, 1$ and all i. In words, the causal IV affects the outcome only through affecting the treatment. Because of the exclusion restriction, we will index potential outcomes in terms of the treatment only so that, for example $Y_i(1) = Y_i(z, d = 1)$ is the outcome that subject i would experience if she were assigned level 1 of the treatment.

CA3-1 *Effective randomness of the* IV: $\{Y(0), Y(1), D(0), D(1)\} \perp\!\!\!\perp Z$, that is, the causal IV Z does not share common causes with the outcome Y and the treatment received D.

2.2 Core Assumptions for Z Being an Intensity Preserving Proxy for a Causal IV Z^*

In some settings, the measured IV Z does not have a causal effect on the treatment itself, but is instead a proxy for an IV Z^* that has a causal effect on the treatment (Hernán and Robins, 2006). In the NICU study from Section 1, Z is excess travel time under average traffic conditions; Z^* might be the actual excess travel time the mother faces at the time she is ready to go to the hospital. When Z does not have a causal effect on the treatment itself, but is instead a proxy for a causal IV Z^*, Hernán and Robins (2006) call Z a surrogate IV and Z^* the causal IV (Note: Hernán and Robins, 2006 use the notation U^* instead of Z^* for the causal IV).

Let $D_i(z^*)$ be the treatment that subject i would receive if she were assigned level z^* of Z^*. Each subject has a set of potential treatments, $\{D_i(z^*), z^* \in \mathcal{Z}^*\}$ where $\mathcal{Z}^*$ is the set of possible values of Z^*.

We assume that the measured IV Z is an intensity preserving proxy for Z^*:

DEFINITION 1. Z is an intensity preserving proxy for $Z^* \in \mathbb{R}$ when:

(i) $F(z^*|Z=0) \geq F(z^*|Z=1)$ for all $z^* \in \mathbb{R}$ and $F(z^*|Z=0) > F(z^*|Z=1)$ for at least one $z^* \in \mathbb{R}$, where F denotes the c.d.f.; and

(ii) $Z \perp\!\!\!\perp \{\{D(z^*), z^* \in \mathcal{Z}^*\}, Y(0), Y(1)\}|Z^*$, where $Y(d)$ is the outcome the subject would experience if her treatment level was set to d.

(i) says that the conditional distribution of $Z^*|Z=1$ strictly stochastically dominates the conditional distribution of $Z^*|Z=0$. (ii) says that Z has no predictive power for the potential treatment received or potential outcomes once we control for Z^*. Note that since D is a function of Z^* and $\{D(z^*), z^* \in \mathcal{Z}^*\}$, (ii) implies that $Z \perp\!\!\!\perp \{\{D(z^*), z^* \in \mathcal{Z}^*\}, D, Y(0), Y(1)\}|Z^*$.

The following are three sets of sufficient conditions for Z to be an intensity preserving proxy for Z^*. First, the intensity preserving proxy property is reflexive in that if $Z = Z^*$, then Z is an intensity preserving proxy. Second, if Z^* is binary, then Z is an intensity preserving proxy if (a) $P(Z^*=1) > P(Z^*=1|Z=0)$ and (b) Z misclassifies Z^* nondifferentially with respect to the treatment and potential outcomes, that is, $P(Z^*=1|Z=z, Y(0)=y_0, Y(1)=y_1, D=d) = P(Z^*=1|Z=z)$ for all z, y_0, y_1, d. Third, if $Z = I(Z^* + W > 0)$ where W has a density function that is log concave and $W \perp\!\!\!\perp \{Z^*, D, Y(0), Y(1)\}$, then Z is an intensity preserving proxy for Z^*. This follows from Lehmann (1966), Example 12; see the supplementary materials for details and also Chalak (2017). Examples of log concave density functions include the normal, uniform, logistic and exponential densities (Bagnoli and Bergstrom, 2005).

The following are "core" assumptions for Z^* to be a valid causal IV (Hernán and Robins, 2006):

CA1-2 *Positive Causal Effect of Z^* on Treatment*: The probability of taking the treatment is a strictly increasing function of Z^*: $E[D(z^*=b)] > E[D(z^*=a)]$ for all $b > a$, $a \in \mathcal{Z}^*$, $b \in \mathcal{Z}^*$;

CA2-2 *Exclusion Restriction*: Let $Y_i(z^*, d)$ be the potential outcome that subject i would experience if the causal IV is set to level z^* and the treatment is set to level d. The exclusion restriction is that for all units i, $Y_i(z^*=a, d) = Y_i(z^*=b, d)$ for all $a \in \mathcal{Z}^*$, $b \in \mathcal{Z}^*$ for $d = 0$ or 1. As discussed in Section 2.1, because of the exclusion restriction, we will index potential outcomes in terms of the treatment only, so that $Y_i(d) = Y_i(z^*, d)$.

CA3-2 *Effective randomness of the IV*: $\{Y(0), Y(1), \{D(z^*), z^* \in \mathcal{Z}^*\}\} \perp\!\!\!\perp Z^*$, that is, the causal IV Z^* does not share common causes with the outcome Y and the treatment received D.

For a causal IV Z^* that satisfies the core assumptions (CA1-2)–(CA3-2) and an intensity preserving proxy Z for Z^*, the following holds (proofs in supplementary materials):

CA-Proxy-Implication-1: $E(D|Z=1) > E(D|Z=0)$.

CA-Proxy-Implication-2: $Y(0), Y(1) \perp\!\!\!\perp Z$.

2.3 Core Assumptions for Z to Be a Valid IV When Neither Z Nor Something That Z Is a Proxy for Can Be Manipulated

In some settings, the IV or something that the IV is a proxy for cannot easily be thought of as being manipulated while keeping everything else about the unit fixed. For example, Neal (1997) used whether a student was Catholic as an IV for the effect of attending a Catholic secondary school vs. a public secondary school on educational achievement. For many people, growing up Catholic shapes their identities in ways that are hard to imagine changing while keeping everything else about the person fixed (Cavolina et al., 2000). Consequently, it is difficult to define potential treatment received, $\{D_i(z=1), D_i(z=0)\}$, since this would require manipulating the person i's religion without changing anything else about the person. Instead, we consider Z fixed and non-manipulable for a person, and let $Y_i(0)$ be the outcome that i would have if her treatment was set to 0 and $Y_i(1)$ be the outcome i would have if her treatment was set to 1. We define Z as a valid non-manipulable IV if it is positively associated with the treatment and independent of potential outcomes, that is, Z satisfies the following core assumptions:

CA1-3 *Positive association between Z and the treatment*: $E(D|Z=1) > E(D|Z=0)$.

CA2-3 *IV is independent of potential outcomes*: $\{Y(0), Y(1)\} \perp\!\!\!\perp Z$.

The core assumptions CA1-3 and CA2-3 correspond with the classical econometric view of IVs (Stock, 2001). For the Catholic school example of Neal (1997), the being Catholic IV could fail CA2-3 if being Catholic is directly relevant to the potential educational achievement a person would have if she were to not go (go) to a Catholic school or if being Catholic is associated with a factor that affects potential educational achievement.

2.4 Common Features of the Core Assumptions for the Three Different Types of IVs

For the core assumptions for the three different types of IVs discussed in Sections 2.1–2.3, we have the following common features:

CF-CA-1 $Y_i(d)$ represents the outcome subject i would have if she were to receive level d of the treatment for $d = 0, 1$.

CF-CA-2 The IV Z is positively associated with the treatment, $E(D|Z = 1) > E(D|Z = 0)$.

CF-CA-3 The IV Z is independent of potential outcomes: $\{Y(0), Y(1)\} \perp\!\!\!\perp Z$.

The basic assumptions BA1–BA2 and the core assumptions are not enough to identify causal effects, and some additional assumption(s) is needed (Angrist, Imbens and Rubin, 1996, Hernán and Robins, 2006). We review in the next section one such set of additional assumptions, deterministic monotonicity in the deterministic compliance class framework (Angrist, Imbens and Rubin, 1996).

3. REVIEW OF DETERMINISTIC COMPLIANCE CLASS FRAMEWORK AND IDENTIFICATION RESULTS

3.1 Deterministic Compliance Class Framework and Deterministic Monotonicity Assumption

The deterministic compliance class framework presented in Angrist, Imbens and Rubin (1996) assumes Z is the causal IV and the core assumptions (CA1-1)–(CA3-1) in Section 2.1. Furthermore, the framework assumes a subject's treatment received is a (subject specific) deterministic function of the level of the subject's IV Z and there are no unrepresented versions of the IV,

DCC-IVA1 *There are no unrepresented versions of the IV*. Regardless of how the IV is administered, $D_i(z)$ is the treatment that subject i would receive if given level z of the IV for $z = 0, 1$.

A subject's compliance class C is $C = \text{nt}$ (never taker) if $D(0) = 0, D(1) = 0$; $C = \text{at}$ (always taker) if $D(0) = 1, D(1) = 1$; $C = \text{co}$ (complier) if $D(0) = 0, D(1) = 1$; and $C = \text{de}$ (defier) if $D(0) = 1, D(1) = 0$. The additional assumptions for Z to be a valid IV in the deterministic compliance class framework is the deterministic monotonicity assumption:

DCC-IVA2 *Deterministic Monotonicity*. $D_i(1) \geq D_i(0)$ for all subjects i, that is, there are no defiers.

3.2 Identification Results Under the Deterministic Compliance Class Framework

Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) showed that under (BA1)–(BA2), (CA1-1)–(CA3-1) and (DCC-IVA1)–(DCC-IVA2), the LATE, $E[Y(1) - Y(0)|C = \text{co}]$, is identified:

$$E\big[Y(1) - Y(0)|C = \text{co}\big] = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{P(D = 1|Z = 1) - P(D = 1|Z = 0)}. \tag{1}$$

The denominator of (1) is the proportion of compliers, $P(C = co) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$.

The average treatment effect for compliers (1) can be estimated by the sample analogue of (1), which is called the Wald or two-stage least squares estimator:

$$\frac{\hat{E}(Y|Z = 1) - \hat{E}(Y|Z = 0)}{\hat{P}(D = 1|Z = 1) - \hat{P}(D = 1|Z = 0)}. \tag{2}$$

3.3 Deterministic Monotonicity with a Proxy IV

We consider that Z is an intensity preserving proxy for a causal IV Z^* that satisfies (CA1-2)–(CA3-2) as described in Section 2.2. Following Hernán and Robins (2006), suppose that the causal IV Z^* has no unrepresented versions and follows a deterministic monotonicity assumption:

DCC-Proxy-IVA1 *There are no unrepresented versions of the causal IV* Z^*. $D_i(z^*)$ is the treatment that subject i would receive if she were given level z^* of the IV regardless of how the IV is administered.

DCC-Proxy-IVA2 *Deterministic Monotonicity of the causal IV* Z^*. $D_i(z^*) \geq D_i(z^{*\prime})$ for all $z^* > z^{*\prime}$ for all subjects i.

Hernán and Robins (2006, Theorem 5) shows that for an intensity preserving proxy Z for a binary causal IV Z^* satisfying (CA1-2)–(CA3-2) and (DCC-Proxy-IVA1)–(DCC-Proxy-IVA2), the right hand side of (1) (the Wald estimand) is equal to the average treatment effect for compliers with respect to the IV Z^*, that is, $E[Y(1) - Y(0)|D(z^* = 1) = 1, D(z^* = 0) = 0]$. In the supplemental materials (Section H), we review results of Hernán and Robins (2006) and Chalak (2017) that show how for a continuous causal IV Z^* satisfying DCC-Proxy-IVA1 and DCC-Proxy-IVA2, the Wald estimand is a weighted average of treatment effects with subjects whose treatment probability changes more given $Z = 1$ vs. $Z = 0$ receiving larger weight.

4. STOCHASTIC COMPLIANCE CLASS FRAMEWORK AND IDENTIFICATION RESULTS

4.1 Stochastic Compliance Class Framework

In the stochastic compliance class framework, we will introduce in this section, we do not assume that a subject's compliance behavior (treatment taken as the level of the IV varies) is deterministic or even that it is well defined. We allow for the IV to have unrepresented versions, that is, violate DCC-IVA1 (or DCC-Proxy-IVA1). We also allow for the IV to be an intensity preserving proxy for a causal IV (Section 2.2) or even for the IV to be non-manipulable (Section 2.3). Furthermore, we allow for the IV to violate deterministic monotonicity as long as it satisfies a weaker stochastic monotonicity condition.

We assume that (BA1)–(BA2) hold as well as the one of the IV frameworks in Sections 2.1–2.3 holds so that the common implications (CF-CA-1)–(CF-CA-3) hold. In order to explain the additional assumptions needed to identify a causal effect in the stochastic compliance class framework, we define $\mathbf{U}$ to be a sufficient set of unmeasured common causes of D and Y if conditional on $\mathbf{U}$ and Z, the effect of D on Y is unconfounded, meaning

$$\{Y(0), Y(1)\} \perp\!\!\!\perp D | Z, \mathbf{U} \tag{3}$$

(VanderWeele and Shpitser, 2013). We always have that $\mathbf{U} = \{Y(0), Y(1)\}$ satisfies (3) but there may be additional $\mathbf{U}$ that satisfy (3), for example, under the deterministic compliance class model, $\mathbf{U} = \{D(0), D(1)\}$ satisfies (3). See Section 6 below for discussion of the choice of $\mathbf{U}$.

The additional assumptions for Z to be a valid IV in the stochastic compliance class framework are that there exists a sufficient set of unmeasured common causes $\mathbf{U}$ [i.e., $\mathbf{U}$ satisfying (3)] such that:

SCC-IVA1 *IV is jointly independent of the potential outcomes and* $\mathbf{U}$: $\{Y(0), Y(1), \mathbf{U}\} \perp\!\!\!\perp Z$.

SCC-IVA2 *Stochastic Monotonicity*: $P(D = 1|Z = 1, \mathbf{U} = \mathbf{u}) \geq P(D = 1|Z = 0, \mathbf{U} = \mathbf{u})$ for all $\mathbf{u}$. This means that the probability of having the treatment is at least as high for subjects with the encouraging level of the IV compared to the non-encouraging level of the IV within all strata of $\mathbf{U}$.

Regarding SCC-IVA1, note that for an IV satisfying one of the sets of core assumptions in Sections 2.1–2.3, so that CF-CA-3 holds which means $\{Y(0), Y(1)\} \perp\!\!\!\perp Z$, the role of furthermore having $\{Y(0), Y(1), \mathbf{U}\} \perp\!\!\!\perp Z$ for identifying treatment effects in the stochastic compliance class framework is similar to the role of having the joint independence $\{Y(0), Y(1), D(0), D(1)\} \perp\!\!\!\perp Z$ for identifying treatment effects in the deterministic compliance class framework. As noted above, there could be more than one sufficient set of unmeasured common causes $\mathbf{U}$ of D and Y satisfying (3) (VanderWeele and Shpitser, 2013), but we say that Z is a valid IV in the stochastic compliance class framework if it satisfies (SCC-IVA1)–(SCC-IVA2) (in addition to BA1, BA2, CF-CA-1, CF-CA-2 and CF-CA-3) for any sufficient set of unmeasured common causes $\mathbf{U}$. If Z is a valid IV in the stochastic compliance class framework, then the causal effect we shall define in Section 4.2 is the same for all sufficient sets of unmeasured common causes $\mathbf{U}$ for which (SCC-IVA1)–(SCC-IVA2) are satisfied.

4.2 Identification Results Under the Stochastic Compliance Class Framework

Let $\mathcal{Q}$ denote the weighted distribution from the population with the weight proportional to

$$w(\mathbf{u}) = P(D = 1|Z = 1, \mathbf{U} = \mathbf{u}) - P(D = 1|Z = 0, \mathbf{U} = \mathbf{u})$$

for a unit with $\mathbf{U} = \mathbf{u}$. The distribution $\mathcal{Q}$ samples more heavily from strata of $\mathbf{U}$ in which the IV is more associated with treatment. Since $\mathcal{Q}$ weights each subject by how strongly the IV is associated with the treatment in that subject's subgroup (which is defined by $\mathbf{U}$), we call the average treatment effect under $\mathcal{Q}$, the *Strength-of-*IV *Weighted Average Treatment Effect (SIV-WATE)*:

$$E_{\mathcal{Q}}[Y(1) - Y(0)] = \frac{\int E[Y(1) - Y(0)|\mathbf{U} = \mathbf{u}] w(\mathbf{u})\, dF(\mathbf{u})}{\int w(\mathbf{u})\, dF(\mathbf{u})}. \tag{4}$$

The following proposition and corollary show that functions of potential outcomes under the weighted distribution $\mathcal{Q}$, in particular the SIV-WATE, are identified by a valid IV under the stochastic compliance class framework.

PROPOSITION 1. *Assume* BA1, BA2, CF-CA-1, CF-CA-2, CF-CA-3, SCC-IVA1 *and* SCC-IVA2 *hold for a* $\mathbf{U}$ *that satisfies* (3). *For any measurable function* g *with* $E|g(Y(1))| < \infty$,

$$E_{\mathcal{Q}}[g(Y(1))] = \frac{E(Dg(Y)|Z = 1) - E(Dg(Y)|Z = 0)}{P(D = 1|Z = 1) - P(D = 1|Z = 0)}, \tag{5}$$

and for any measurable function g with $E|g(Y(0))| < \infty$,

$$
\begin{aligned}
&E_{\mathcal{Q}}\big[g(Y(0))\big] \\
&\quad = -\big(E\big((1-D)g(Y)|Z=1\big) \\
&\qquad - E\big((1-D)g(Y)|Z=0\big)\big) \\
&\qquad /\big(P(D=1|Z=1) - P(D=1|Z=0)\big).
\end{aligned} \tag{6}
$$

As a result,

$$
\begin{aligned}
&E_{\mathcal{Q}}\big[g(Y(1)) - g(Y(0))\big] \\
&\quad = \frac{E(g(Y)|Z=1) - E(g(Y)|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}.
\end{aligned} \tag{7}
$$

COROLLARY 1. *Assume* BA1, BA2, CF-CA-1, CF-CA-2, CF-CA-3, SCC-IVA1 *and* SCC-IVA2 *hold for a* $\mathbf{U}$ *that satisfies* (3). *Then the* SIV-WATE, $E_{\mathcal{Q}}[Y(1) - Y(0)]$, *equals*

$$
\begin{aligned}
&E_{\mathcal{Q}}\big[Y(1) - Y(0)\big] \\
&\quad = \frac{E(Y|Z=1) - E(Y|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}.
\end{aligned} \tag{8}
$$

The right-hand side of (8) is the probability limit of the Wald estimator (2). Thus, Corollary 1 shows that if Z is a valid IV under the stochastic compliance class framework and we use the usual Wald (two stage least squares) estimator, then we obtain a consistent estimate of the SIV-WATE.

In the supplementary materials (Section D), we show that the deterministic compliance class results reviewed in Section 3 are a special case of the stochastic compliance class framework identification results of this section.

4.3 No Sign Reversal Property Under Stochastic Monotonicity

When treatment effects are heterogeneous, Imbens and Angrist (1994) showed that the probability limit of the Wald estimator (2) has a disturbing sign reversal property: it is possible for the treatment effect to be positive for every subject but for the Wald estimator to converge in probability to a negative number. However, under deterministic monotonicity, if the sign of the treatment effects ($+, 0$ or $-$) is the same for every subject in the population, then the sign of the treatment effects is identified because the sign of the probability limit of the Wald estimator (2) is equal to the average treatment effect for compliers. Corollary 1 shows that this no sign reversal property also holds under stochastic monotonicity: if the sign of the treatment effects ($+$, 0 or $-$) is the same for every subject in the population, then the sign of the treatment effects is identified because the identified SIV-WATE is a weighted average of treatment effects.

4.4 Characterizing the Strength of IV Weighted Population $\mathcal{Q}$ in Terms of Observed Covariates

The SIV-WATE is the average treatment effect for the weighed population $\mathcal{Q}$. To understand $\mathcal{Q}$ better, it is useful to characterize how the distribution of the observed covariates for $\mathcal{Q}$ relates to that of the unweighted population, for example, compare $E_{\mathcal{Q}}[A]$ to $E[A]$ for an observed covariate A.

PROPOSITION 2. *Assume that* BA1, BA2, CF-CA-1, CF-CA-2, CF-CA-3, SCC-IVA1 *and* SCC-IVA2 *hold for a* $\mathbf{U}$ *that satisfies* (3) *and that the following extended versions of* SCC-IVA1 *and* (3) *involving* A *hold*:

$$
\big\{Y(0), Y(1), A\big\} \perp\!\!\!\perp D|Z, \mathbf{U}, \tag{9}
$$

$$
\big\{Y(0), Y(1), \mathbf{U}, A\big\} \perp\!\!\!\perp Z. \tag{10}
$$

Then,

$$
\begin{aligned}
&E_{\mathcal{Q}}[A] \\
&\quad = \frac{E(DA|Z=1) - E(DA|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}.
\end{aligned} \tag{11}
$$

For an IV that is effectively randomly assigned, (10) will hold. For potential choices of $\mathbf{U}$ for which (9) and SCC-IVA2 may hold, see Section 6. Note that if we condition on an observed covariate A as we discuss in Section 7, then (9) will automatically hold for this A. Proposition 2 is a generalization of results for deterministic monotonicity that characterize the compliers in terms of their distribution of observed covariates (Angrist and Pischke, 2009).

4.5 Bounds on the Global Average Treatment Effect

Under the stochastic monotonicity assumption, a valid IV identifies a weighted average of treatment effects, the SIV-WATE (Corollary 1). The IV does not identify the unweighted, global average treatment effect, $E[Y(1) - Y(0)]$, but if a researcher is able to put bounds on how much the average treatment effect varies as $\mathbf{U}$ varies, that is, denoting the range $\sup_{\mathbf{u}} E[Y(1) - Y(0)|\mathbf{U} = \mathbf{u}] - \inf_{\mathbf{u}} E[Y(1) - Y(0)|\mathbf{U} = \mathbf{u}]$ by $\text{range}_{\mathbf{u}}\text{ATE}$, the researcher puts a bound

$$
\text{range}_{\mathbf{u}}\text{ATE} \leq r, \tag{12}
$$

then knowing the SIV-WATE will bounds the global average treatment effect.

PROPOSITION 3. *Suppose* (12) *holds for some positive* r *and that* BA1-BA2, (CF-CA-1)–(CF-CA-3) *and* (SCC-IVA1)–(SCC-IVA2) *hold. Then, the following are bounds on* $E[Y(1) - Y(0)]$:

$$a - rb \leq E\big[Y(1) - Y(0)\big] \leq a + rb, \tag{13}$$

where

$$a = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)},$$
$$b = 1 - \big[E(D|Z=1) - E(D|Z=0)\big].$$

When there are no bounds or other constraints on the outcome besides (12), the bounds in Proposition 3 are sharp because they are attained for distributions of the potential outcomes that are consistent with the observable data and the deterministic compliance class model holding; see the discussion below the proof of Proposition 8 (the analogue of Proposition 3 with covariates) in the supplemental materials. This also shows that when there are no bounds or other constraints on the outcome besides (12), the bounds under deterministic monotonicity and stochastic monotonicity are the same. When there are bounds on the outcome, for example, the outcome is binary, then the bounds in Proposition 3 can potentially be tightened. Section E of the supplementary materials presents an algorithm for finding the bounds for a binary outcome by extending the approach of Ramsahai (2012). For a binary outcome, bounds under stochastic monotonicity can be wider than under deterministic monotonicity; see Section E of the supplementary materials.

The bounds from Proposition 3 can be considerably tighter than the bounds for an IV that does not satisfy stochastic monotonicity. Section F of the supplementary materials provides an example.

5. SENSITIVITY ANALYSIS FOR VIOLATIONS OF STOCHASTIC MONOTONICITY

Regardless of whether stochastic monotonicity holds, as long as the other conditions hold for Z to be a valid IV in the stochastic compliance class framework (BA1, BA2, CF-CA-1, CF-CA-2, CF-CA-3 and SCC-IVA1), the quantity on the right-hand side of (7) with $g(Y) = Y$ that we use to estimate the SIV-WATE is equal to

$$\begin{aligned}&\frac{E(Y|Z=1) - E(Y|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)} \\ &\quad = \frac{E[E(Y(1) - Y(0)|\mathbf{U})w(\mathbf{U})]}{E[w(\mathbf{U})]}\end{aligned} \tag{14}$$

since the proof of Proposition 1 does not make use of stochastic monotonicity. When stochastic monotonicity is violated, (14) is not a weighted average of treatment effects because some of the "weights" $w(\mathbf{U})$ are negative. In this case, we might be interested in the strength of IV weighted average of treatment effects among subjects for whom the weights $w(\mathbf{U})$ are positive, which we call the positive strength of IV weighted average treatment effect (PSIV-WATE). The PSIV-WATE is equal to the following, where we let $\mathcal{A} = \{\mathbf{U} : w(\mathbf{U}) \geq 0\}$,

$$\begin{aligned}&\text{PSIV-WATE} \\ &= E_{\mathcal{Q}}\big[Y(1) - Y(0)|\mathbf{U} \in \mathcal{A}\big] \\ &= E\bigg[E\big[Y(1) - Y(0)|\mathbf{U}\big]\frac{w(\mathbf{U})1\{\mathbf{U} \in \mathcal{A}\}}{E[w(\mathbf{U})1\{\mathbf{U} \in \mathcal{A}\}]}\bigg],\end{aligned} \tag{15}$$

where $1\{\cdot\}$ denotes the indicator function. When stochastic monotonicity holds, the PSIV-WATE is the SIV-WATE and equals the right hand side of (7) with $g(Y) = Y$. When stochastic monotonicity does not hold, then the following theorem gives the asymptotic bias from using the sample analogue of the right-hand side of (7) with $g(Y) = Y$ to estimate the PSIV-WATE, where we define the negative strength of IV weighted average treatment effect (NSIV-WATE) as the weighted average treatment effect among subjects for whom the $w(\mathbf{U})$ are negative and the subjects are weighted by the absolute value of $w(\mathbf{U})$,

$$\begin{aligned}&\text{NSIV-WATE} \\ &= E_{\mathcal{Q}}\big[Y(1) - Y(0)|\mathbf{U} \in \mathcal{A}^C\big] \\ &= E\bigg[E\big[Y(1) - Y(0)|\mathbf{U}\big]\frac{w(\mathbf{U})1\{\mathbf{U} \in \mathcal{A}^C\}}{E[w(\mathbf{U})1\{\mathbf{U} \in \mathcal{A}^C\}]}\bigg].\end{aligned} \tag{16}$$

PROPOSITION 4. *When* BA1, BA2, CF-CA-1, CF-CA-2, CF-CA-3 *and* SCC-IVA1 *hold but the stochastic monotonicity condition* SCC-IVA2 *may not hold,*

$$\begin{aligned}&\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} - \text{PSIV-WATE} \\ &\quad = -\lambda(\text{NSIV-WATE} - \text{PSIV} - \text{WATE}),\end{aligned} \tag{17}$$

where

$$\lambda = -\frac{E[w(\mathbf{U})1\{\mathbf{U} \in \mathcal{A}^C\}]}{E[w(\mathbf{U})]}.$$

Proposition 4 generalizes the formula in Angrist, Imbens and Rubin (1996), Proposition 3, for the bias from using the Wald estimate to estimate the LATE when there are defiers; when the deterministic compliance

class framework holds, then (17) is equal to the bias formula in Angrist, Imbens and Rubin (1996). The bias due to violations of stochastic monotonicity is composed of two factors. The first factor λ is related to the proportion of subjects for whom stochastic monotonicity is violated and is equal to zero under the stochastic monotonicity assumption. The numerator of λ relates to the proportion of subjects for whom stochastic monotonicity is violated and the magnitude by which stochastic monotonicity is violated for these subjects—the smaller this proportion and magnitude, the smaller the numerator will be. The denominator of λ is equal to the overall strength of the association between the IV and the treatment,

$$\begin{aligned} E\big[w(\mathbf{U})\big] &= E\big[P(D=1|Z=1,\mathbf{U}) \\ &\qquad - P(D=1|Z=0,\mathbf{U})\big] \\ &= P(D=1|Z=1) - P(D=1|Z=0) \end{aligned}$$

[see (41) in the supplemental materials]. The stronger the IV is, the less sensitive the IV estimate is to violations of stochastic monotonicity. The second factor in the bias formula, PSIV-WATE–NSIV-WATE, is related to the difference in treatment effects between those subjects for whom treatment is positively associated with the IV and those subjects for whom treatment is negatively associated with the IV. The less difference there is between treatment effects for these two types of subjects, the less bias there is from violations of stochastic monotonicity.

6. CHOICE OF U FOR INTERPRETING THE SIV-WATE

Corollary 1 shows for any $\mathbf{U}$ satisfying (3) such that the IV Z satisfies (SCC-IVA1)–(SCC-IVA2) for this $\mathbf{U}$, the probability limit of the Wald estimator (2) is equal to a weighted average of treatment effects, where the weight for the subgroup of units with $\mathbf{U}=\mathbf{u}$ depends on the size of the subgroup and how strongly the IV is associated with the treatment among units in the subgroup. There may be multiple $\mathbf{U}$'s that satisfy (3) such that Z satisfies (SCC-IVA1)–(SCC-IVA2), and thus Corollary 1 may provide multiple interpretations of what the Wald estimator is estimating. We now discuss various possible choices of $\mathbf{U}$.

One choice of $\mathbf{U}$ is $\mathbf{U}=\{Y(0),Y(1)\}$. This $\mathbf{U}$ always satisfies (3) and (SCC-IVA1) simplifies to the IV being independent of the potential outcomes $\{Y(0),Y(1)\}$, which is core assumption CF-CA-3. However, it may not be easy to think about whether stochastic monotonicity holds for this $\mathbf{U}$ because this $\mathbf{U}$ is not tied closely to D. Similarly, interpreting the SIV-WATE in terms of this $\mathbf{U}$ may be difficult because the values of the weight

$$\begin{aligned} &w\big(y(0)=a,y(1)=b\big) \\ &\quad = P\big(D=1|Z=1,Y(0)=a,Y(1)=b\big) \\ &\qquad - P\big(D=1|Z=0,Y(0)=a,Y(1)=b\big) \end{aligned}$$

may be hard to think about since $\{Y(0),Y(1)\}$ are not tied closely to D. A further drawback to $\mathbf{U}=\{Y(0),Y(1)\}$ is that the bounds given by Proposition 3 are the weakest possible; for example, if m_1 is the maximum possible value of Y and m_0 is the minimum possible value and if $\{Y(1)=m_1,Y(0)=m_0\}$, $\{Y(1)=m_0,Y(0)=m_1\}$ both have positive probability mass or probability density, then $\text{range}_{\mathbf{u}}ATE$ is equal to twice the range of Y.

When the deterministic compliance class framework holds so that $\{D(0),D(1)\}$ are well defined, the choice of $\mathbf{U}=\{D(0),D(1)\}$ leads to the usual interpretation of the Wald estimator as estimating the complier average causal effect when deterministic monotonicity holds. This $\mathbf{U}$ always satisfies (3) because $\{D(0),D(1),Z\}$ together determine D under the deterministic compliance class framework. Condition (SCC-IVA1) requires that the IV be independent not only of the potential outcomes $\{Y(0),Y(1)\}$ but also of the potential treatment received $\{D(0),D(1)\}$. The stochastic monotonicity condition for $\mathbf{U}=\{D(0),D(1)\}$ is equal to the deterministic monotonicity condition DCC-IVA2 that there are no defiers. An advantage of the choice of $\mathbf{U}=\{D(0),D(1)\}$ is that it is relatively easy to think about whether stochastic monotonicity holds since one just has to think about, is there anybody who would do the opposite of what the IV encourages? Similarly, interpreting the SIV-WATE for this $\mathbf{U}$ is relatively easy since the weights are 1 and 0, and the SIV-WATE is just the average treatment effect for compliers. Another advantage of the choice of $\mathbf{U}=\{D(0),D(1)\}$ compared to $\mathbf{U}=\{Y(0),Y(1)\}$ is that applying Proposition 3 to $\mathbf{U}=\{D(0),D(1)\}$ may yield tighter bounds, especially if D and $Y(0)$, $Y(1)$ are thought to be weakly correlated. However, if we do not have a good understanding of the characteristics of compliers vs. non-compliers, then we may not feel comfortable choosing $\text{range}_{\mathbf{u}}ATE$ to be that much less than twice the range of Y and the bounds from Proposition 3 will not be that much tighter than for $\mathbf{U}=\{Y(0),Y(1)\}$. Furthermore, if we do not have a good understanding of the characteristics of compliers vs. non-compliers, even though we know the SIV-WATE is the average treatment effect for compliers, it

will be hard to interpret the type of person the SIV-WATE most applies to.

As discussed in the introduction, $\{D(0), D(1)\}$ are sometimes not well defined because there are unrepresented versions of the IV as in the NICU study and the deterministic compliance class framework does not hold. For such settings, a possible choice of $\mathbf{U}$ is the average value that D would take over all versions of the IV. For example, in the NICU study, we could let $\mathbf{U}_i$ be the chance that mother i would deliver at a high level NICU if she were assigned to live in a random zip code with probability proportional to the number of deliveries in the zip code. The stochastic monotonicity condition here is that for each level of $\mathbf{U}$, the chance of delivering at a high level NICU is at least as great when averaging over zip codes with excess travel time ≤ 10 minutes as when averaging over zip codes with excess travel time > 10 minutes (where both averages are weighted by the number of deliveries in the zip code). Stochastic monotonicity allows for the possibility of "defier" zip codes like in Table 1 as long as the "complier" zip codes outweigh the "defier" zip codes. In order to use this $\mathbf{U}$ (average value that D would take over all versions of the IV) to interpret the SIV-WATE using the results in Section 4, in addition to stochastic monotonicity (SCC-IVA2) holding, we need SCC-IVA1 and (3) to hold. If the IV is effectively randomly assigned, then SCC-IVA1 will hold. For (3) to hold, a sufficient condition is that for any characteristic that is associated with the potential outcomes, among subjects whose proportions of $D = 1$ would be the same over randomly assigned versions of the IV, the proportions of $D = 1$ would be the same over randomly assigned versions of the $Z = 1$ level of the IV for all strata of the characteristic. For example, consider the characteristic of a mother's economic situation. If, among mothers who would go to a high level NICU $x\%$ of the time for any fixed x, poor mothers are equally likely to go to a high level NICU when living far away from one than not poor mothers, then (3) will be satisfied. But if among these mothers who would go to a high level NICU say 50% of the time, poor mothers are less likely to go to a high level NICU when living far away from one than not poor mothers and also correspondingly less likely to go to a high level NICU when living close to one, then (3) will be violated. In this case we could append $\mathbf{U}$ with the mother's economic situation and all other characteristics that are associated with the potential outcomes and for which among subjects whose proportion of $D = 1$ would be the same over randomly assigned versions of the IV, the proportions would be different over randomly assigned versions of only the $Z = 1$ level of the IV, and then (3) will be satisfied for the appended $\mathbf{U}$. Then, as long as the appended $\mathbf{U}$ still satisfies SCC-IVA1 and stochastic monotonicity SCC-IVA2, the appended $\mathbf{U}$ can be used to interpret the SIV-WATE.

The choice of $\mathbf{U}$ as the average value of D over the different versions of the IV is an analogue to $\mathbf{U}$ being the compliance class $[\mathbf{U} = \{D(0), D(1)\}]$ that allows for the deterministic compliance class framework to not hold and has similar characteristics. With the choice of $\mathbf{U}$ as the average value that D would take over the different versions of the IV, we can interpret the SIV-WATE as a weighted average of treatment effects that puts more weight on subjects whose treatment choice is more influenced by the IV and applying Proposition 3 with this $\mathbf{U}$ may yield tighter bounds than with $\mathbf{U} = \{Y(0), Y(1)\}$, especially if D and $\{Y(0), Y(1)\}$ are thought to be weakly correlated.

Other choices of $\mathbf{U}$ that are in between $\{Y(0), Y(1)\}$ and $\{D(0), D(1)\}$ but keep D and $\{Y(0), Y(1)\}$ conditionally independent can be considered. Section 9 considers such a choice for a physician prescribing preference IV. Typically, choosing $\mathbf{U}$ to be something that is as closely correlated to D as possible but still satisfies stochastic monotonicity will lead to the tightest bounds on the average treatment effect using Proposition 3. Also, ideally, $\mathbf{U}$ would represent something that *in principle* can be measured, so that we can know how much the SIV-WATE weights a particular subject. For example, if the deterministic compliance class model and deterministic monotonicity hold, and a subject knows herself well enough to know her compliance class, then she knows whether the SIV-WATE applies to her (if she is a complier) or does not (if she is a never taker or always taker). Similarly, if there are different versions of the IV and the deterministic compliance class model does not hold, but the IV satisfies the stochastic compliance class model assumptions SCC-IVA1–SCC-IVA2 with $\mathbf{U}$ being the average value of D over the different versions of the IV as in the above paragraph, then a subject who knows she is likely to take the treatment when assigned a version of the IV with $Z = 1$ but not likely to take the treatment when assigned a version of the IV with $Z = 0$ understands that subjects like her are weighted heavily in the SIV-WATE whereas a subject who is likely to take the treatment whether assigned a version of the IV with $Z = 1$ or $Z = 0$ understands that subjects like her are weighted less in the SIV-WATE.

7. CONDITIONING ON COVARIATES

In observational studies, a potential IV Z might only be independent of potential outcomes and unmeasured common causes of D and Y after conditioning on certain measured covariates $\mathbf{X}$. For example, in the NICU study, race is associated with the proposed IV, excess travel time, and race is also thought to be associated with infant mortality through mechanisms other than NICU level that are not fully measured in our data such as previous cesarean section, inadequate prenatal care, and chronic maternal medical conditions (Lorch et al., 2012b). Consequently, excess travel time is only plausibly independent of potential outcomes and unmeasured common causes of D and Y after conditioning on $\mathbf{X}$ = race. The IV method can still be used to learn about a weighted average of treatment effects as long as Z is a valid IV under the stochastic compliance class framework within each strata of $\mathbf{X}$, that is, Z satisfies BA1, BA2, CF-CA-1, a conditional version of CF-CA-2 that $E(D|Z=1,\mathbf{X}) \geq E(D|Z=0,\mathbf{X})$ for all $\mathbf{X}$ with strict inequality for at least one $\mathbf{X}$, a conditional version of CF-CA-3 and SCC-IVA1 that $\{Y(0), Y(1), \mathbf{U}\} \perp\!\!\!\perp Z|\mathbf{X}$ and a conditional version of SCC-IVA2 that $E(D|Z=1,\mathbf{X},\mathbf{U}) \geq E(D|Z=0,\mathbf{X},\mathbf{U})$. This extension is formulated and analogues of Propositions 1, 2, 3 and 4 are proved in the supplemental materials (Sections A and B). Such results are comparable to various previous results under the deterministic monotonicity assumption (Abadie, 2003, Tan, 2006, Ogburn, Rotnitzky and Robins, 2015).

8. APPLICATION TO STUDY THE EFFECTIVENESS OF HIGH-LEVEL NEONATAL INTENSIVE CARE UNITS

We consider the study of the effect on mortality for premature babies of being delivered in a high level vs. low level NICU discussed in the introduction. The data is from Pennsylvania from 1995–2005 (192,078 premature babies); see Lorch et al. (2012a) for full description. The data was collected from birth and death certificates and the UB-92 form that hospitals use for billing purposes. A baby's health status before delivery is an important confounder as mothers are more likely to go to a high level NICU if a baby is considered to be at high risk for complications or death. The data contains some measures of the baby's health such as gestational age, but the data is also missing several important measures available to the doctor and mother when deciding where to deliver such as fetal heart tracing results, the severity of maternal problems during pregnancy (e.g., the data contains an indicator for whether a mother had pregnancy-induced hypertension but no information on the severity) and the mother's adherence to prenatal guidelines. Concern about these unmeasured confounders motivated Baiocchi et al. (2010), Lorch et al. (2012a), Yang, Lorch and Small (2014) and Guo et al. (2014) to use an IV approach. We follow Yang, Lorch and Small (2014) in considering the IV Z to be whether or not the mother's excess travel time from the nearest high level NICU compared to the nearest low level NICU is less than or equal to 10 minutes ($Z=1$ vs. $Z=0$). The travel time is computed using Dijkstra's (Dijkstra, 1959) algorithm for the shortest path between the centroid of the mother's zip code and the hospital under average traffic conditions as implemented in ArcView software.

As discussed in the introduction, deterministic monotonicity is not plausible for the excess travel time IV because excess travel time is determined by the zip code a mother lives in and other characteristics of the zip code influence hospital choices (e.g., community, family and friends' views about the different hospitals in the area) such that there is more encouragement to go to high level NICUs in certain zip codes which are far from high level NICUs than in certain zip codes which are close to high level NICUs (see Table 1). The excess travel time IV is also likely to violate deterministic monotonicity because the travel time computed from the ArcView software is the travel time under average traffic conditions and may not accurately represent the traffic conditions faced by a mother at the time when she needs to decide where to deliver; also, if the mother uses public transportation, the travel time depends on public transportation routes. Although deterministic monotonicity is not plausible, stochastic monotonicity is plausible. Consider Z_i^* to be the actual excess travel time mother i faces at the time she is ready to go to the hospital. Z is plausibly an intensity preserving proxy for Z^* since actual excess travel times are likely to be longer for mothers whose ArcView measured travel times are greater than 10 minutes ($Z=1$) than for mothers whose ArcView measured travel time is $\leq$10 minutes ($Z=0$), and the actual NICU delivered at and potential outcomes presumably would not depend on Z if we knew the actual excess travel time Z^*. Consider $\mathbf{U}_i$ to be the chance that mother i would deliver at a high level NICU if she were assigned to live in a random zip code with probability proportional to the number of deliveries in the zip code. Z plausibly satisfies the stochastic compliance class IV assumptions with this $\mathbf{U}$ for the following reasons:

- CA1-2: A mother typically obtains prenatal care from and would prefer to deliver at a close by hospital so that a smaller excess travel time to the nearest high level NICU makes a mother more likely to deliver at a high level NICU (Phibbs et al., 1993).
- CA2-2: Most mothers have time to reach either the nearest high level or low level NICU before delivering so that the marginal travel time should not directly affect outcomes (Lorch et al., 2012a).
- (3): As discussed in Section 6, a sufficient condition for (3) to hold for this $\mathbf{U}$ (proportion of $D = 1$ over randomly assigned versions of the IV) is that conditional on measured characteristics $\mathbf{X}$, for any unmeasured characteristic C that is associated with potential outcomes, among subjects whose proportions of $D = 1$ over randomly assigned versions of the IV (zip codes, weighted by deliveries) is the same, the proportions of $D = 1$ over randomly assigned $Z = 1$ versions of the IV (zip codes with excess travel time $\leq$10, weighted by deliveries) would also be the same for all strata of the unmeasured characteristic C. We are not aware of a reason to expect this sufficient condition to be violated for the NICU study but also do not have any supportive evidence for the condition holding.
- CA3-2 and SCC-IVA1: Women do not expect to have a premature delivery, and thus conditional on measured socioeconomic variables such as mother's education and measured zip code characteristics such as average income levels and poverty rates, women do not choose where to live based on distance to a high level NICU, making independence of excess travel time from potential outcomes and $\mathbf{U}_i$ plausible (Lorch et al., 2012a).
- SCC-IVA2. Within each strata of mother's general tendency to deliver at a high level NICU, it is plausible that differences between the ArcView travel time and actual travel time, and differences in factors like community, family and friends' beliefs about the hospitals in the area average out between the $Z = 1$ and $Z = 0$ mothers so that within each strata of $\{\mathbf{U}, \mathbf{X}\}$, the mothers with $Z = 1$ (ArcView excess travel time $\leq$ 10 minutes) are more likely to go to a high level NICU than mothers with $Z = 0$ (ArcView excess travel time $>$ 10 minutes).

We estimated the SIV-WATE by (24) in the supplementary materials using logistic regression to estimate $E(Y|Z, \mathbf{X})$ and $P(D|Z, \mathbf{X})$; we also estimated the SIV-WATE for three ranges of gestational ages—moderate to late preterm (33–37 weeks), very preterm (28–32 weeks) and extremely preterm ($\leq$27 weeks) based on (25) in the supplemental materials. Table 2 shows the estimates. The estimates are expressed in terms of the effect of delivering at a high level NICU vs. a low level NICU on mortality per 1000 births. 95% confidence intervals were computed using the percentile bootstrap, with the resampling stratified on the three ranges of gestational ages. The SIV-WATE estimate is that being delivered in a high level NICU prevents 6 deaths per 1000 births with a 95% confidence interval of preventing 2.5 to 8.7 deaths; our analysis suggests that high level NICUs are effective for the SIV-WATE population. The effect of high level NICUs on reducing mortality is estimated to be greater for more premature babies, with a particularly large effect for extremely premature babies ($\leq$27 weeks).

TABLE 2
SIV-WATE estimates and confidence intervals for effect of delivering in high level NICUs vs. low level NICUs on mortality per 1000 *premature births*

Group	Estimate	95% CI
All	−6.0	(−8.7, −2.5)
Gestational Age, 33–37 wks	−1.8	(−2.8, −0.7)
Gestational Age, 28–32 wks	−26.0	(−38.5, −11.0)
Gestational Age, ≤27 wks	−110.7	(−164.8, −45.4)

Table 3 compares the distribution of characteristics A in the strength of IV weighted population $\mathcal{Q}$ to the unweighted population using the analogue of Proposition 2 (Proposition 7 in the supplemental materials) The strength-of-IV weighted population is similar to the full population in terms of mother's education, race and comorbidities, but not in terms of gestational age. The strength of IV weighted population has more moderate to late premature babies (33–37 weeks) and less very (28–32 weeks) or extremely ($\leq$27 weeks) premature babies. Since the effect of high level NICUs appears to be greater among very and extremely premature babies (Table 2), the SIV-WATE for all babies may underestimate the global average treatment effect for all babies.

We now consider sensitivity to violations of stochastic monotonicity using Proposition 9 in the supplemental materials, the analogue of Proposition 4 with covariates. Proposition 9 provides information about how much bias there is in the estimates in Table 2 as estimates of the PSIV-WATE, the strength of IV weighted average treatment effect for subgroups of subjects whose chance of delivering at a high level

TABLE 3
Characteristics of SIV-WATE weighted population compared to unweighted population for NICU study

Characteristic X	Prevalence of X in weighted population $\mathcal{Q}$	Prevalence of X in unweighted population	Ratio of Prevalence in $\mathcal{Q}$ to unweighted population
Gestational age, 33–37 wks	0.90	0.87	1.03
Gestational age, 28–32 wks	0.08	0.10	0.84
Gestational age, ≤27 wks	0.02	0.03	0.58
Birthweight <1500 g	0.06	0.09	0.73
Mother College Graduate	0.26	0.26	0.99
African American	0.16	0.16	0.96
Gestational Diabetes	0.05	0.05	1.00
Diabetes mellitus	0.02	0.02	0.92
Pregancy-induced hypertension	0.10	0.10	0.95
Chronic hypertension	0.02	0.02	0.93

NICU is positively affected by living near to a high level NICU. We consider, is it plausible that the PSIV-WATE is zero or positive (high level NICUs are equivalent or harmful compared to low level NICUS) even though Table 2 suggests high level NICUs are beneficial for the PSIV-WATE population? To use Proposition 9 for a sensitivity analysis, we need to choose a value(s) of λ. To do so interpretably, we can estimate the denominator of λ (which is equal to $E[P(D=1|Z=1,\mathbf{X}) - P(D=1|Z=0,\mathbf{X})]$) by $\frac{1}{N}\sum_{i=1}^{N} \hat{P}(D=1|Z=1,\mathbf{X}) - \hat{P}(D=1|Z=0,\mathbf{X})$, which equals 0.40 for the data set. One way of interpreting this quantity is that for a randomly drawn baby and a random draw from that baby's compliance distribution, the difference between the probability of being a complier and being a defier is 0.40. The numerator of λ can be interpreted as the weighted average of the probability of being a defier minus the probability of being a complier among those strata of $\mathbf{U}$ for which there are more defiers than compliers, weighted by the size of the strata, times the probability of being in a strata of $\mathbf{U}$ for which there are more defiers than compliers; an upper bound on this numerator is the probability of being in a strata of $\mathbf{U}$ for which there are more defiers than compliers. Denote the numerator of λ by ξ. Suppose that the difference between the PSIV-WATE and the NSIV-WATE is at most 24.2 deaths per 1000 births; this number is chosen because it is the difference between the estimated SIV-WATEs for gestational age of 33–37 weeks and gestational age of 28–32 weeks in Table 2, two fairly different risk groups. Then, by Proposition 4, the estimated overall PSIV-WATE could range from $-6 - (\xi/0.4)24.2$ to $-6 + (\xi/0.4)24.2$. The upper bound of the range is below zero for $\xi < 0.099$ and the upper 95% confidence bound $(-2.5 + (\xi/0.4)24.2)$ is below zero for $\xi < 0.041$. We consider a departure of stochastic monotonicity of these magnitudes to be large as we estimated the departure from stochastic monotonicity if $\mathbf{U}$ were equal to the observed covariates but we had not controlled for them to be 0.029 (we estimated this by fitting logistic regressions of D on $\mathbf{X}$ for the $Z=1$ subjects and D on $\mathbf{X}$ for the $Z=0$ subjects and then applying the estimates to all subjects, seeing for what proportion of subjects, the latter estimate was higher). Thus, the inference that delivering at a high-level NICU reduces mortality for the PSIV-WATE population is robust to at least a moderate departure from stochastic monotonicity.

We now consider putting bounds on the global average treatment effect using Proposition 8 in the supplemental materials, the analogue of Proposition 3 with covariates. Suppose that for fixed $\mathbf{X}$ and varying $\mathbf{U}$, $E[Y(1) - Y(0)|\mathbf{U}=\mathbf{u},\mathbf{X}=\mathbf{x}]$ has a range from $(1/m) \times E_{\mathcal{Q}}[Y(1) - Y(0)|\mathbf{X}]$ to $m \times E_{\mathcal{Q}}[Y(1) - Y(0)|\mathbf{X}]$ depending on sensitivity parameter m so that $r = (m - \frac{1}{m}) \times E_{\mathcal{Q}}[Y(1) - Y(0)|\mathbf{X}]$ in Proposition 8. Then, using Proposition 8, we estimate the bounds on $E[Y(1) - Y(0)|\mathbf{X}]$ to be

$$\hat{E}_{\mathcal{Q}}\big[Y(1) - Y(0)|\mathbf{X}\big]\big(1 \pm \big\{(m - 1/m)\big[\hat{P}(D=1|Z=1,\mathbf{X}) - \hat{P}(D=1|Z=0,\mathbf{X}) - 1\big]\big\}\big).$$

We have $E[Y(1) - Y(0)] = E[E[Y(1) - Y(0)|\mathbf{X}]]$, which we can estimate by $\frac{1}{N}\sum_{i=1}^{N} \hat{E}[Y(1) - Y(0)|\mathbf{X}_i]$. Substituting the estimated lower and upper bounds into this latter expression provides estimates of the

TABLE 4
Bounds on global average treatment effect (mortality per 1000 births) when for fixed $\mathbf{X}$ *and varying* $\mathbf{U}$, $E[Y(1) - Y(0)|\mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}]$ *has a range of* $(m - \frac{1}{m}) \times E_{\mathcal{Q}}[Y(1) - Y(0)|\mathbf{X}]$

m	Estimated Bounds	95% CI for bounds
1.1	(−9.5, −8.6)	(−15.2, −3.7)
1.5	(−11.0, −7.1)	(−17.7, −3.0)
2	(−12.6, −5.5)	(−20.1, −2.3)
3	(−15.3, −2.8)	(−24.8, −1.2)
5	(−20.4, 2.2)	(−32.9, 3.6)

bounds on $E[Y(1) - Y(0)]$. Table 4 shows the estimated bounds for $m = 1.1, 1.5, 2, 3, 5$ along with 95% confidence bounds formed by bootstrap resampling and using the Bonferroni method, which means taking the 2.5th percentile of the bootstrapped lower bound and 97.5th percentile of the upper bound (Horowitz and Manski, 2000, Cheng and Small, 2006). For moderate amounts of treatment effect heterogeneity, $m = 1.1$ to 3, the upper bound is below zero so that there is evidence that delivering all premature babies at high level NICUs compared to delivering all premature babies at low level NICUs would reduce mortality.

9. PHYSICIAN PRESCRIBING PREFERENCE IV

For comparing the effectiveness or safety of two drugs, physician's prescribing preference has often been used as an IV (Brookhart et al., 2006). Boef et al. (2016) and Swanson et al. (2015) present evidence that deterministic monotonicity can be violated for physician prescribing preference IVs. Suppose Z_i is whether the ith patient's physician last prescribed the treatment or control, and that Z_i is an intensity preserving proxy for the proportion of patients that patient i's physician would prescribe the treatment to, which we denote by Z_i^*. The deterministic monotonicity assumption for Z_i^*, DCC-Proxy-IVA2, implies that if a physician with $Z^* = z^*$ would give the patient the treatment, then all physicians with $Z^* > z^*$ would also give the patient the treatment (Hernán and Robins, 2006). Swanson et al. (2015) considered prescription of atypical versus conventional antipsychotic medication in the elderly. Based on a survey of physicians' preferences for treating different types of patients, they found that deterministic monotonicity was violated—for 85% of the patients, there was at least one physician who would prescribe conventional antipsychotic medication and another physician who would prescribe atypical antipsychotic medication even though the second physician had greater preference for conventional antipsychotic medication.

We now consider stochastic monotonicity for the physician prescribing preference IV. Consider $\mathbf{U}$ to be the vector of all patient characteristics that systematically affect treatment or potential outcomes so that $D \perp\!\!\!\perp \{Y(0), Y(1)\}|\mathbf{U}$; patients who share the same $\mathbf{U}$ can be considered patients of the same type. Consider the following assignment process for a patient's physician and the last patient seen by the physician: a patient's physician is chosen randomly and the last patient the physician saw before the current patient is chosen randomly among all other patients. Such an assignment process is plausible when the physicians are at a single practice and which physician a patient sees is essentially chosen randomly (Korn and Baumrind, 1998). Under this assignment process, (3) is satisfied, and also since the IV is randomly assigned, the IV is independent of $\{Y(1), Y(0), \mathbf{U}\}$, meaning SCC-IVA1 is satisfied. The stochastic monotonicity condition, SCC-IVA2, says that for each patient type, the chance that a patient of that type will receive the treatment if the physician's previous patient received the treatment is at least as large as if the physician's previous patient received the control. Suppose that the assignment process described above holds where there are J types of patients with probabilities p_j (so that the type of the previous patient seen by a physician is multinomial with probabilities $p_1, \ldots, p_J$) and a physician always makes the same prescription to the same type of patient. Let A_{jk} denote whether the kth physician would prescribe the treatment to a patient of type j and let there be K physicians. Then the stochastic monotonicity condition is that for all patient types $j = 1, \ldots, J$,

$$\frac{\sum_{k=1}^{K}\sum_{j'=1}^{J} A_{jk}p_{j'}A_{j'k}}{\sum_{k=1}^{K}\sum_{j'=1}^{J} p_{j'}A_{j'k}} \geq \frac{\sum_{k=1}^{K}\sum_{j'=1}^{J} A_{jk}p_{j'}(1 - A_{j'k})}{\sum_{k=1}^{K}\sum_{j'=1}^{J} p_{j'}(1 - A_{j'k})}. \tag{18}$$

Boef et al. (2016) found that in a survey of physicians' preferences for treating patients with subclinical hypothyroidism, preferences for starting vs. not starting treatment on levothryoxine, deterministic monotonicity was violated but the stochastic monotonicity condition (18) held. See the supplemental materials (Sections I and J) for two other examples in which deterministic monotonicity is likely violated but stochastic monotonicity is plausible.

10. DISCUSSION

Angrist, Imbens and Rubin's (1996) work on the LATE has had a large influence on how researchers understand and interpret what is learned from IV analyses. However, there has been controversy over whether the LATE is a useful estimand because (i) the LATE is an average treatment effect over a subpopulation, the compliers, that cannot be identified in the sense that there are subjects for whom we do not know whether they belong to the subpopulation and (ii) the LATE may not be the treatment effect of primary interest, instead the global average treatment effect is often of greater interest (Deaton, 2010, Pearl, 2011, Swanson and Hernán, 2014). Our generalization of the LATE, the SIV-WATE, could be criticized along the same lines. We support Imbens' discussion (Section 4.6 of Imbens, 2014) of why the LATE is useful and we feel similar reasoning also applies to the SIV-WATE. For studying a treatment, we would ideally like to have a randomized trial with perfect compliance. IV analysis is only used when for practical or ethical reasons, we do not have this ideal study and instead have an observational study with unmeasured confounding (or a randomized trial with noncompliance); as Imbens says, "IV analysis is an analysis in a second-best setting." Under deterministic monotonicity, the LATE tells us what we can learn directly about treatment effects from the data without making homogeneity assumptions about the treatment effect. Under stochastic monotonicity, the SIV-WATE tells us what we can learn directly about treatment effects from the data without making either homogeneity assumptions about the treatment effect or the assumption of deterministic monotonicity. We can describe the weighted population that the SIV-WATE refers to in terms of the observed covariates, as we did in the NICU study (Table 3). The SIV-WATE can be combined with assumptions about how heterogeneous treatment effects are to find bounds on the global average treatment effect (Section 4.5 and Table 4).

The SIV-WATE is also useful in some settings because it directly addresses the decisionmaking question of interest. Suppose a doctor is trying to decide whether to encourage a patient to take a treatment with a possible side effect. Suppose the patient's utility for taking the treatment is $y - s1(\text{side effect occurs})$ where the side effect occurs with probability t if the treatment is taken. The difference in the expected utility from encouraging the patient to take the treatment vs. not is

$$\begin{aligned}\int &\big[P(D=1|Z=1,\mathbf{U}=\mathbf{u})\\ &\quad - P(D=1|Z=0,\mathbf{U}=\mathbf{u})\big]\{E\big[Y(1)\\ &\quad - Y(0)|\mathbf{U}=\mathbf{u}\big] - st\}\,dF(\mathbf{u})\\ &= (\text{SIV-WATE} - st) \times \int w(\mathbf{u})\,dF(\mathbf{u}).\end{aligned}$$

Therefore, the doctor would like to encourage the patient to take the treatment if the SIV-WATE is greater than st.

ACKNOWLEDGMENTS

We thank the Associate Editor, Editor and referees for constructive comments that helped improve our paper.

SUPPLEMENTARY MATERIAL

Supplement to "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption" (DOI: 10.1214/17-STS623SUPP; .pdf). Section A of the supplementary materials presents analogues of Propositions 1–4 when conditioning on observed covariates. Section B presents proofs for the propositions given in Section A of the supplementary materials. Section C presents proofs for the results in Section 2 of the main text. Section D shows that deterministic compliance class framework identification results are a special case of stochastic compliance class framework results. Section E discusses bounds on the global average treatment effect for a binary outcome under stochastic monotonicity. Section F gives an example in which bounds under stochastic monotonicity are tighter than bounds without stochastic monotonicity. Section G extends results in the main text to the setting of a non-binary instrumental variable. Section H discusses identification results when the measured IV is an intensity preserving proxy for a continuous causal IV satisfying deterministic monotonicity. Sections I and J give additional examples besides those given in the main text in which deterministic monotonicity is likely violated but stochastic monotonicity is plausible.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ANGRIST, J. D. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

BAGNOLI, M. and BERGSTROM, T. (2005). Log-concave probability and its applications. *Econom. Theory* **26** 445–469. MR2213177

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects for studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

BOEF, A. G. C., GUSSEKLOO, J., DEKKERS, O. M., FREY, P., KEARNEY, P. M., KERSE, N., MALLEN, C. D., MCCARTHY, V. J. C., MOOIJAART, S. P., MUTH, C., RODONDI, N., ROSEMANN, T., RUSSELL, A., SCHERS, H., VIRGINI, V., DE WAAL, M. W. M., WARNER, A., LE CESSIE, S. and DEN ELZEN, W. P. J. (2016). Physician's prescribing preference as an instrumental variable: Exploring assumptions using survey data. *Epidemiology* **27** 276–283.

BROOKHART, M. A. and SCHNEEWEISS, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *Int. J. Biostat.* **3** Art. 14, 25. MR2383610

BROOKHART, M. A., WANG, P., SOLOMON, D. H. and SCHNEEWEISS, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17** 268–275.

CAVOLINA, M. J. F., KELLY, M. A. T., STONE, J. A. J. and DAVIS, R. G. M. (2000). *Growing up Catholic*: *The Millennium Edition*: *An Infinitely Funny Guide for the Faithful*, *the Fallen and Everyone in-Between*. Image.

CHALAK, K. (2017). Instrumental variables methods with heterogeneity and mismeasured instruments. *Econometric Theory* **33** 69–104. MR3574861

CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 815–836. MR2301296

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York. MR0095561

DE CHAISEMARTIN, C. (2017). Tolerating defiance: Local average treatment effects without monotonicity. *Quant. Econ.* **8** 367–396.

DEATON, A. (2010). Instruments, randomization and learning about development. *J. Econ. Lit.* **48** 424–455.

DIJKSTRA, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math.* **1** 269–271. MR0107609

DINARDO, J. and LEE, D. S. (2011). Program evaluation and research designs. In *Handbook of Labor Economics*, *Vol.* 4 463–536. Elsevier, Amsterdam.

FINKELSTEIN, A., TAUBMAN, S., WRIGHT, B., BERNSTEIN, M., GRUBER, J., NEWHOUSE, J. P., ALLEN, H., BAICKER, K. and GROUP OREGON HEALTH STUDY (2012). The Oregon health insurance experiment: Evidence from the first year. *Q. J. Econ.* **127** 1057–1106.

GUO, Z., CHENG, J., LORCH, S. A. and SMALL, D. S. (2014). Using an instrumental variable to test for unmeasured confounding. *Stat. Med.* **33** 3528–3546. MR3260644

HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372.

HOROWITZ, J. L. and MANSKI, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Amer. Statist. Assoc.* **95** 77–88. MR1803142

HUBER, M. and MELLACE, G. (2012). Relaxing monotonicity in the identification of local average treatment effects. Working paper.

IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **61** 467–476.

KLEIN, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *J. Econometrics* **155** 99–116. MR2607188

KORN, E. L. and BAUMRIND, S. (1998). Clinician preferences and the estimation of causal treatment differences. *Statist. Sci.* **13** 209–235. MR1665709

LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153. MR0202228

LORCH, S. A., BAIOCCHI, M., AHLBERG, C. E. and SMALL, D. S. (2012a). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics* **130** 270–278.

LORCH, S. A., KROELINGER, C. D., AHLBERG, C. and BARFIELD, W. D. (2012b). Factors that mediate racial/ethnic disparities in US fetal death rates. *Am. J. Publ. Health* **102** 1902–1910.

MANSKI, C. F. (1990). Non-parametric bounds on treatment effects. *Am. Econ. Rev.* **80** 351–374.

MCCLELLAN, M., MCNEIL, B. J. and NEWHOUSE, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *J. Am. Med. Dir. Assoc.* **272** 859–866.

NEAL, D. (1997). The effects of Catholic secondary schooling on educational achievement. *J. Labor Econ.* **14** 98–123.

OGBURN, E. L., ROTNITZKY, A. and ROBINS, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 373–396. MR3310531

PEARL, J. (2011). Principal stratification—a goal or a tool? *Int. J. Biostat.* **7** Art. 20, 15. MR2787410

PHIBBS, C. S., MARK, D. H., LUFT, H. S., PELTZMAN-RENNIE, D. J., GARNICK, D. W., LICHTENBERG, E. and MCPHEE, S. J. (1993). Choice of hospital for delivery: A comparison of high-risk and low-risk women. *Health Serv. Res.* **28** 201.

RAMSAHAI, R. R. (2012). Causal bounds and observable constraints for non-deterministic models. *J. Mach. Learn. Res.* **13** 829–848. MR2913720

ROBINS, J. 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, eds.) 113–159.

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

RUBIN, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

SMALL, D. S., TAN, Z., RAMSAHAI, R. R., LORCH, S. A. and BROOKHART, M. A. (2017). Supplement to "Instrumental variable estimation with a stochastic monotonicity assumption." DOI:10.1214/17-STS623SUPP.

STOCK, J. H. (2001). Instrumental variables in economics and statistics. In *International Encyclopedia of the Social & Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.) 7577–7582. Elsevier, Amsterdam.

SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. MR3264549

SWANSON, S. A., MILLER, M., ROBINS, J. M. and HERNÁN, M. A. (2015). Definition and evaluation of the monotonicity condition for preference-based instruments. *Epidemiology* **26** 414–420.

TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *J. Amer. Statist. Assoc.* **101** 1607–1618. MR2279483

TAN, Z. (2010). Marginal and nested structural models using instrumental variables. *J. Amer. Statist. Assoc.* **105** 157–169. MR2757199

VANDERWEELE, T. J. and SHPITSER, I. (2013). On the definition of a confounder. *Ann. Statist.* **41** 196–220. MR3059415

YANG, F., LORCH, S. A. and SMALL, D. S. (2014). Estimation of causal effects using instrumental variables with nonignorable missing covariates: Application to effect of type of delivery NICU on premature infants. *Ann. Appl. Stat.* **8** 48–73. MR3191982