Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Nonparametric Bounds in Two-Sample Summary-Data Mendelian Randomization: Some Cautionary Tales for Practice

Ralph Møller Trane | Hyunseung Kang

[1]Department of Statistics, University of Wisconsin–Madison

Correspondence
Ralph Møller Trane, Department of Statistics, University of Wisconsin–Madison.
Email: rtrane@wisc.edu

Present Address
1300 University Avenue, Madison, WI 53706

Recently, in genetic epidemiology, Mendelian randomization (MR) has become a popular approach to estimate causal effects by using single nucleotide polymorphisms from genome-wide association studies (GWAS) as instruments. The most popular type of MR study, a two-sample summary-data MR study, relies on having summary statistics from two independent GWAS and using parametric methods for estimation. However, little is understood about using a nonparametric bound-based analysis, a popular approach in traditional instrumental variables frameworks, to study causal effects in two-sample MR. In this work, we explore using a bound-based analysis in two-sample MR studies, focusing primarily on implications for practice. We also propose a framework to assess how likely one can obtain more informative bounds if we used a different MR design, notably a one-sample MR design. We conclude by demonstrating our findings through two real data analyses concerning the causal effect of smoking on lung cancer and the causal effect of high cholesterol on heart attacks. Overall, our results suggest that while a bound-based analysis may be appealing due to its nonparametric nature, it is rarely suitable in practice as a method of analysis to bound the causal exposure effect in two-sample MR studies unless strong assumptions are met.

KEYWORDS:
Causal Inference, Nonparametric bounds, Mendelian randomization

1 | INTRODUCTION

In recent years, genetic variants have been used as instrumental variables (IV) to estimate causal effects in epidemiological studies, often referred to as Mendelian randomization (MR) studies [1,2,3]. Typically, MR studies are based on a two-sample study design where published summary statistics from two independent genome wide association studies (GWAS), with one providing information about the exposure and instrument and the other about the outcome and instrument, are used [4,5,6]. Under a two-sample study design, investigators frequently use parametric methods to study exposure effects. Examples include the IVW estimator [4], MR-Egger regression [7], weighted median [8], MR-PRESSO [9] and MR-RAPS [10], to name a few; see [3,11,12] for recent reviews.

An alternative approach to study exposure effects in instrumental variables without parametric assumptions is through nonparametric IV bounds [13,14,15,16,17]. Briefly, nonparametric IV bounds use a minimum set of assumptions to provide a range of plausible values for the exposure effect. They are typically used when the outcome, the exposure, and the instrument are all binary and are simultaneously observed; we refer to this setting as the one-sample setting to contrast it from the two-sample setting. Arguably, the most well-known IV bounds are the Balke-Pearl bounds [13] for the average treatment effect. [14] and [16] extended

the Balke-Pearl bounds to allow for a non-binary instrument.[18] derived bounds under two-sample study designs; see[19] and references therein for a recent summary of IV bounds.

Using IV bounds can be an attractive alternative to study exposure effects in non-MR, one-sample settings[20,19] and some[21,22] have suggested using IV bounds to study exposure effects in MR studies given the strong parametric assumptions accompanying most MR analyses. Despite these suggestions, to the best of our knowledge and compared to parametric methods, there is little work on actually using bounds in typical MR, i.e. two-sample study designs with summary statistics, nor any practical guidance on when a bound-based analysis may be useful to bound the exposure effect. For example, what kind of genetic variants provide the most informative conclusions about the exposure effect in terms of the bounds not containing the null effect? Can combining multiple variants lead to shorter and tighter bounds? How do the bounds change if many instruments are weak, which is typical in MR studies from two-sample designs? The overall goal of this paper is to offer some practical guidance on using IV bounds in two-sample MR studies. We focus on two aspects of bounds that will better inform MR investigators about the exposure effect: (1) the length of the bounds and (2) whether bounds cover the null effect of zero (i.e. direction/sign of the effect).

Our overall takeaway message for investigators using two-sample MR studies is that unless the genetic instruments are strongly associated with the exposure, a bound-based analysis will often be non-informative. In particular, even with more assumptions than the usual set of assumptions for bounds, the width of the two-sample bound is only guaranteed to be less than 1 if the risk difference of the instrument's association to the exposure is greater than 0.5. Also, with summary data from two samples, combining multiple instruments will not yield a shorter bound than simply using the strongest instrument. Finally, investigators should either try to use one-sample MR studies as the bounds can be narrower under one-sample MR studies compared to two-sample MR studies.

2 | METHODS

2.1 | Review: Notation, Definitions, and Assumptions

Let X and Y be binary exposure and outcome variables, respectively, Z be a categorical instrumental variable taking values in $\{0, 1, 2\}$, and U be an unmeasured confounder for the effect of X on Y. Let $Y^{z,x}$ be the potential outcome[23,24] had the subject received exposure value $X = x$ and instrument value $Z = z$. We assume the stable unit treatment value assumption (SUTVA)[25,26], formalized as $Y = \sum_{x,z} I[Z = z, X = x]Y^{x,z}$ where $I[\cdot]$ is the indicator function.

We make the following set of assumptions about X, Y, Z, and U that are found in MR studies; see[21] and[27] for details.

(A1) *(Relevance)*: $Z \not\perp X$
(A2) *(Independent instrument)*: $Z \perp U$
(A3) *(Exclusion restriction)*: $Y^{z,x} = Y^{z',x} = Y^x$ for all x, z, z'
(A4) *(Conditional ignorability of X, Z given U)*: $Y^{z,x} \perp Z, X|U$

Briefly, (A1) can be satisfied by finding SNPs that have been consistently associated with the exposure. (A2) and (A3) are justified by scientific theory and can be violated if the SNP is (i) in linkage disequilibrium with an unmeasured SNP that affects the exposure and the outcome or (ii) has multiple functions beyond affecting the exposure (i.e. pleiotropic), to name a few. Finally, (A4) states that if U is observed, then it is sufficient to unconfound the relationship between X and Y. Throughout the paper, we will assume (A1)-(A4) hold to focus the discussion on the bounds, even though they are important to assess in practice.

We make some brief, additional remarks about assumptions (A1)-(A4). First, in practice, most MR studies only explicitly state assumptions (A1)-(A3) along with some parametric modeling assumptions[3]. Second,[16] showed that one can remove (A4) and strengthen (A2) with $Z \perp U, Y^{z,x}$ without consequence on the IV bounds. Third, under SUTVA and assumptions (A3)-(A4), we have $Y \perp Z|X, U$, which is another common way to express the exclusion restriction in MR studies[21,19]. Fourth, for simplicity, we do not assume the existence of a potential treatment X^z.

Next, we introduce the following assumptions (A5) and (A6); these assumptions are not necessary to construct bounds, but they will help characterize IV bounds in two-sample studies.

(A5) *(Monotonicity between Z and X)* $P(X = 1|Z = z, U) \leq P(X = 1|Z = z + 1, U)$ for $z = 0, 1, 2$

(A6) *(Monotonicity between Z and Y)* $P(Y = 1|Z = z, U) \leq P(Y = 1|Z = z + 1, U)$ for $z = 0, 1, 2$

A variant of (A5) is common in the IV literature to study noncompliance[28,29]. (A6) is an extension of (A5) to the outcome variable. (A5) or (A6) is plausible in MR if the direction of the genetic instrument's effect on the exposure or the outcome is well-established from scientific theory.

We also define instrument strength ST as

$$\text{ST} = \max_{z_1 \neq z_2} |P(X = 1|Z = z_1) - P(X = 1|Z = z_2)| \tag{1}$$

ST reduces to the definition of instrument strength in[13] when the instrument is binary;[13] used ST to characterize the width of their IV bounds. Also, (1) differs from other definitions of instrument strength based on a parametric model between the exposure and the outcome, say the concentration parameter; see[30] for an overview.

Another popular summary statistic measuring instrument strength in MR studies is the coefficient from a logistic model[2,31,32,33,34]. We will utilize a logistic model for the exposure, where the log-odds is given as a linear combination of an intercept, the instruments, and an unmeasured confounder U from the standard normal, and a logistic model for the outcome, where the log-odds is given as a linear combination of an intercept, the exposure, and the unmeasured confounder U. Specifically, we will assume that $P(Z = 0) = P(Z = 2) = 0.25$ and $P(Z = 1) = 0.5$, and a value of an unmeasured confounder U from the standard normal. We will assume the exposure X is binary with

$$\text{logit}(P(X = 1|Z_1 = z_1, ..., Z_p = z_p, U = u)) = \gamma_0 + \sum_i \gamma_i z_i + \gamma_U u, \tag{2}$$

where $\text{logit}(a) = \frac{1}{1+\exp(a)}$ and γ_i corresponds to the estimand of the regression estimate one would obtain from GWAS studying the relationship between the genetic variant and the exposure. This model has been used in MR studies by[31] and[35] so that every instrument estimates the same exposure effect. Similarly, we will assume that the outcome Y is binary with $\text{logit}(P(Y = 1|X = x, U = u)) = \beta_0 + \beta_X \cdot x + \beta_U \cdot u$, which we use to compute the true ATE.

2.2 | IV Bounds Under Two-Sample Designs and Goals of Paper

The most popular design in MR studies is a two-sample design which has two separate data sources, one providing information about (X, Z) in the form of $P(X = 1|Z = z)$, $z \in \{0, 1, 2\}$, and another providing information about (Y, Z) in the form of $P(Y = 1|Z = z)$, $z \in \{0, 1, 2\}$. A two-sample design differs from a more traditional one-sample design which has a single data source providing information on all observed variables (X, Y, Z) in the form of $P(Y = y, X = x|Z = z)$ or related parametrizations. IV bounds have been well-studied in one-sample designs and there is a rich array of guidance for practitioners on how to use them in their own studies[13,16,19]. However, as noted in the introduction, not much is known about the behavior of IV bounds under a two-sample design, especially when and how MR investigators should use them in their own studies.

Formally, the goal of this paper is to offer useful practical advice on using IV bounds to study the average treatment effect (ATE), defined as

$$\text{ATE} = E[Y^1 - Y^0] = \int P(Y = 1 \mid X = 1, U = u)P(U = u)du - \int P(Y = 1 \mid X = 0, U = u)P(U = u)du$$

based on using $P(Y = 1|Z = z)$ and $P(X = 1|Z = z)$ for each $z = 0, 1, 2$ obtained from a two-sample design. Specifically, under a two-sample design and assumptions (A1)-(A4),[18] derived a sharp bound for the ATE (see Appendix A for a detailed review and a discussion of "IV Inequalities"[13,36] in two-sample MR studies).

$$\max\left\{\begin{array}{ll}\max_{z_1\neq z_2} & P(Y=1|Z=z_1)-2\cdot P(Y=1|Z=z_2)-2\cdot P(X=1|Z=z_2)\\ \max_{z_1\neq z_2} & P(Y=1|Z=z_1)+P(X=1|Z=z_1)-P(Y=1|Z=z_2)-P(X=1|Z=z_2)-1\\ \max_{z_1\neq z_2} & 2\cdot P(Y=1|Z=z_1)+2\cdot P(X=1|Z=z_1)-P(Y=1|Z=z_2)-3\\ \max_{z} & -P(Y=1|Z=z)-P(X=1|Z=z)\\ \max_{z} & P(Y=1|Z=z)+P(X=1|Z=z)-2\end{array}\right\}$$

$$\leq ATE \leq \tag{3}$$

$$\min\left\{\begin{array}{ll}\min_{z_1\neq z_2} & P(Y=1|Z=z_1)-2\cdot P(Y=1|Z=z_2)+2\cdot P(X=1|Z=z_2)+1\\ \min_{z_1\neq z_2} & P(Y=1|Z=z_1)+2\cdot P(Y=1|Z=z_2)-2\cdot P(X=1|Z=z_2)+1\\ \min_{z_1\neq z_2} & P(Y=1|Z=z_1)-P(X=1|Z=z_1)+P(X=1|Z=z_2)-P(Y=1|Z=z_2)+1\\ \min_{z} & P(X=1|Z=z)-P(Y=1|Z=z)+1\\ \min_{z} & P(Y=1|Z=z)-P(X=1|Z=z)+1\end{array}\right\}$$

This paper studies two properties of the above bounds that can better guide practice: (1) the length of the bounds and (2) the ability to obtain bounds not covering the null effect of zero. To better understand bound-specific characteristics not due to sampling errors, we will assume we have population-level quantities of $P(Y=1|Z=z)$ and $P(X=1|Z=z)$; in practice, these are estimated summary GWAS statistics from marginal logistic models like the ones introduced in Section 2.1.

3 | PROPERTIES OF IV BOUNDS

3.1 | Length of Bounds and Coverage of Null Effect

Theorem 1 characterizes the length of the IV bound in equation (3) under two-sample designs and assumptions (A1)-(A6); the extra assumptions (A5)-(A6) simplify the formula for the length of the bound to be an interpretable, linear function of instrument strength ST.

Theorem 1. Under assumptions (A1)-(A6), a sharp upper bound on the length of the bound in equation (3) is $2-2\cdot \text{ST}$, i.e. there exists a data generating process satisfying (A1)-(A6) and has width equal to $2-2\cdot \text{ST}$.

See Appendix B for the proof, which extends Theorem 1 to instruments with 2, 3, or 4 categories. Compared to the Balke-Pearl IV bounds with a binary IV in single-sample designs whose width is $1-\text{ST}$[13], the length of the two-sample bounds can be twice as long. Also, the length of two-sample IV bounds is only guaranteed to be less than 1 if instrument strength ST is greater than 0.5; note that this does not imply that instruments with ST less than 0.5 has length greater than 1. In contrast, one-sample IV bounds always have length less than 1 unless ST is zero. In short, there is a cost, in length, of using a two-sample design instead of a one-sample design when performing a bound-based analysis of the ATE in MR.

Figure 1a numerically illustrates the consequences of Theorem 1 by calculating the bounds in equation (3) from 10,000 randomly generated sets of values of $P(X=1|Z=z)$ and $P(Y=1|Z=z)$ that satisfy the IV inequalities and assumptions (A1)-(A4). We also use three real-world data examples where the causal effects are known to exist: the effect of high cholesterol on incidence of heart attacks[37], the effect of smoking on incidence of lung cancer[38], and the effect of obesity on incidence of heart attacks[39]. The first two studies are discussed in detail in Section 5. We see that the width of the bounds often exceed 1 as the instrument strength decreases. Also, the three real-world studies generally do not lead to bounds with length less than 1. Figure 1b further illustrates this point by characterizing the relationship between instrument strength ST and the summary statistic coefficient γ_1 from the logistic exposure model introduced in Section 2.1 with $p=1$; see Appendix C for details. We see that instrument strength ST of 0.5 corresponds to a regression coefficient γ_1 of approximately 1.1, 1.16, 1.4 and 1.8 if γ_U is 0.1, 0.5, 1 and 2, respectively. Coefficients with such magnitudes are rare in GWAS where genetic variants often explain a small amount of variation in the exposure. Also, these values of γ_1 correspond to odds ratios between 3 and 6 and exceed some well-known magnitudes of causal effects in cancer studies, say the effect of exposure to ultraviolet radiation on the incidence of skin cancer where the odds ratios are estimated to be in the range of 1.4 to 2.22[40].

Next, for bounds with length less than 1, we examine what kind of γ_1 is needed in order for the two-sample IV bounds to exclude the null effect of 0 for an anticipated effect size of the ATE. This question is akin to computing the power of bounds but

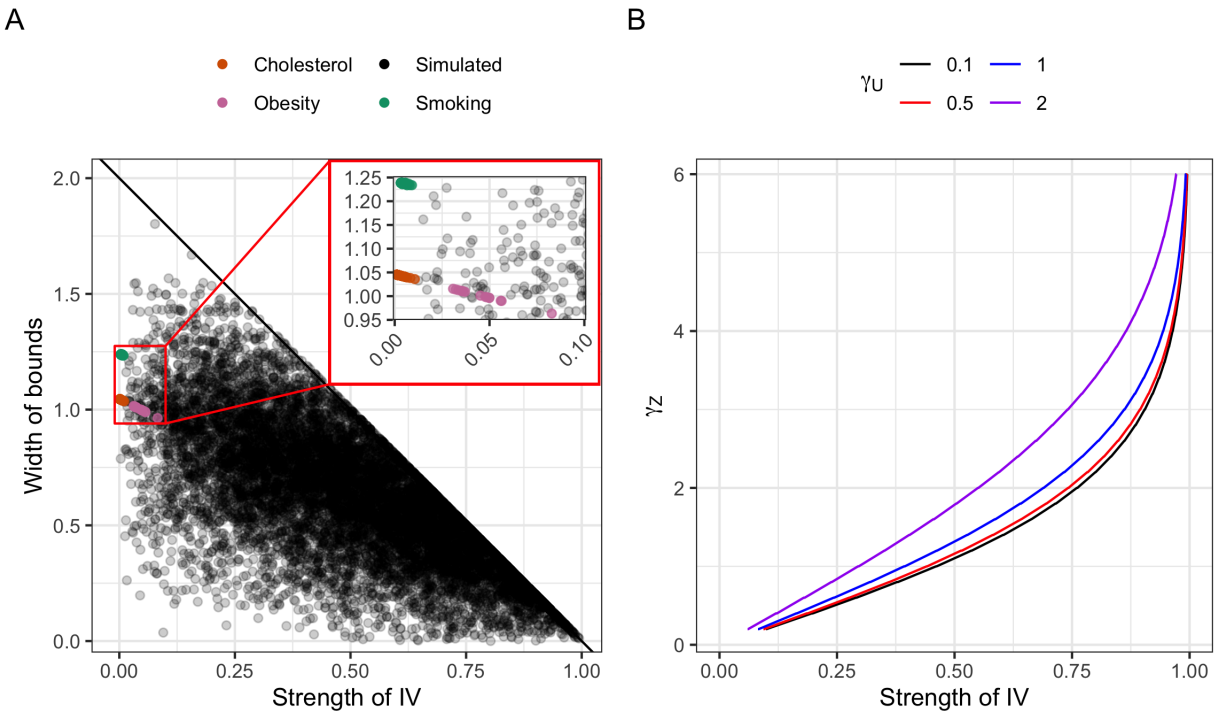

FIGURE 1 Illustration of the relationship between instrument strength, width of bounds obtained from two-sample design, and coefficients from logistic regression model. A: Relationship between instrument strength (ST) and width of the IV bounds. Black line is the upper bound on the two-sample IV bounds from Theorem 1. Black dots indicate one of the 10,000 IV bounds. Colored dots indicate bounds from real data; see Section 5 for details. B: Coefficients from logistic regression model and instrument strength (ST).

with population-level quantities. We reuse the same logistic model above for the exposure and the outcome; see Appendix C for details on this setup. Figure 2 shows the smallest γ_1 needed to exclude 0 for different values of the ATE. Even for moderate effect sizes of 0.4, the corresponding γ_1 must be around 2, a tall order for most GWAS. Also, as the effect of unmeasured confounding increases via γ_U, a larger γ_1 is needed to exclude the null effect.

Overall, using two-sample MR studies with a bound-based analysis is unlikely to be informative. The bounds will often have length greater than 1 and rarely exclude 0 unless very strong genetic variants are used.

3.2 | Would Multiple Instruments Help?

Based on the results above with a single instrument, a natural question from investigators is whether using multiple instruments can lead to more informative bounds for the ATE; see[22] for a recent discussion on this point. For example, suppose we aggregate two-sample IV bounds across multiple instruments by taking intersections of individual IV bounds. This approach may be inferior to another alternative where we expand the levels of Z from $0, 1, 2$ to accommodate multiple instruments[22], but has the benefit of being applicable to most two-sample MR studies with summary statistics from GWAS. However, as we show here, the strongest instrument essentially determines the length of the intersection bound because the bounds from each instrument exhibit a nesting property. That is, using one bound based on the strongest instrument provides the same amount of information about the ATE as the intersection of several bounds from multiple instruments.

We consider the exposure and outcome models introduced in Section 2.1. $z_i \in \{0, 1, 2\}$ represents the ith instrument, and γ_i represents the ith instrument's effect on the exposure. Also, for each instrument i, we set $P(Z_i = 0) = P(Z_i = 2) = 0.25$ and $P(Z_i = 1) = 0.5$. We set $p = 10$ or $p = 50$, and draw U from a standard normal distribution. For simplicity, we set $\beta_U = \gamma_U$, and $\gamma_0 = -\sum_i \gamma_i$ and $\beta_0 = -\beta_1/2$ to spread out the probabilities $P(X = 1|Z = z)$ and $P(Y = 1|X = x)$ as much as possible. β_1 is set to be either 0.25, 0.5, 1, 1.5, or 2. We then consider four scenarios for setting the γ_i's:

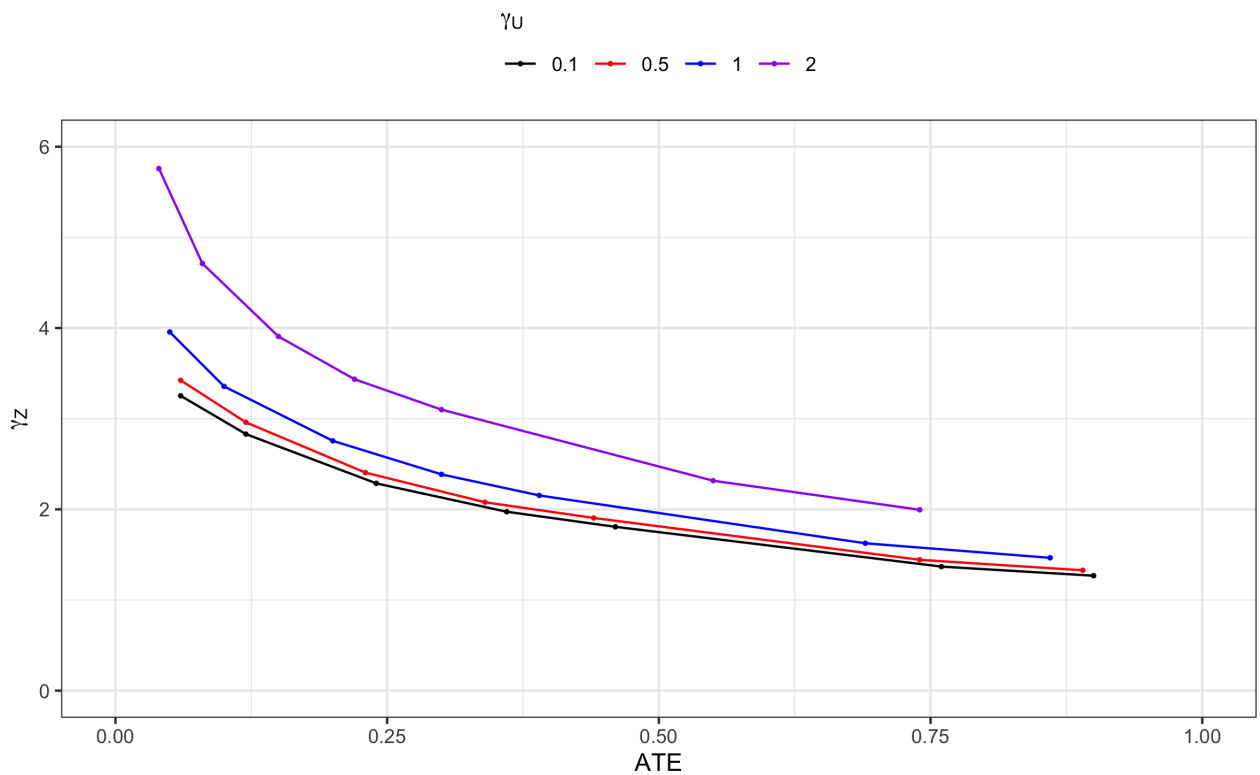

FIGURE 2 The smallest γ_1 needed for a two-sample IV bound to exclude 0.

1. *Many weak instruments*: γ_i are spread out evenly on the interval 0 to 0.2.
2. *Many strong instruments*: γ_i are spread out evenly on the interval 1 to 4. This is the magnitude of γs that detected the direction of the ATE in the previous section
3. *Many very weak instruments, one medium strength instrument*: γ_i, $i = 1, 2, ..., p-1$, are evenly spread out on the interval 0 to 0.01, and $\gamma_p = 0.2$.
4. *Many medium strong instruments, one strong instrument*: γ_i, $i = 1, 2, ..., p-1$, are evenly spread out on the interval 1 to 1.2, and $\gamma_p = 4$.

The first scenario mimics typical magnitudes of coefficients seen in MR studies, and is an example where many genetic traits weakly contribute to the expression of complex traits[41,42,43]. The third scenario represents a genetic architecture where only few genetic variants have strong effects on the exposure while others have weak effects[44]. Scenarios 2 and 4 are as scenarios 1 and 3, but with coefficients of larger magnitude. We don't expect to observe this in practice, but these are the magnitudes that our results in Section 3.1 suggests would result in informative bounds when $p = 1$.

For each scenario, we use monte carlo integration with $1,000,000$ re-samples to obtain $P(X = 1|Z_j = z_j)$ and $P(Y = 1|Z_j = z_j)$ – this procedure is as described in Appendix C, but with multiple instruments. We then use these quantities to obtain two-sample IV bounds for each of the p instruments. Figures 4, 5, 6, and 7 summarize the results. We see that in scenarios 1 and 2, every bound is non-informative, with widths close to or exceeding 1. Also, the bounds are nested within each other. Thus, if we were to aggregate the bounds by taking intersections, the width of the intersection bounds will still be close to or exceed 1. In addition, the increase in magnitude of the γ_i coefficient did not improve the bounds. Scenarios 3 and 4 show similar results in that the bounds cover the null effect, but the strongest instrument in each scenario produces a much smaller bound than in scenarios 1 and 2. From Figure 5 it is clear that on the scale that is often observed in MR studies, two-sample nonparametric bounds are generally non-informative. Also, the bounds in scenarios 3 and 4 are again nested leaving us with the conclusion that the intersection of bounds from multiple instruments will give no more information than the strongest of the instruments itself.

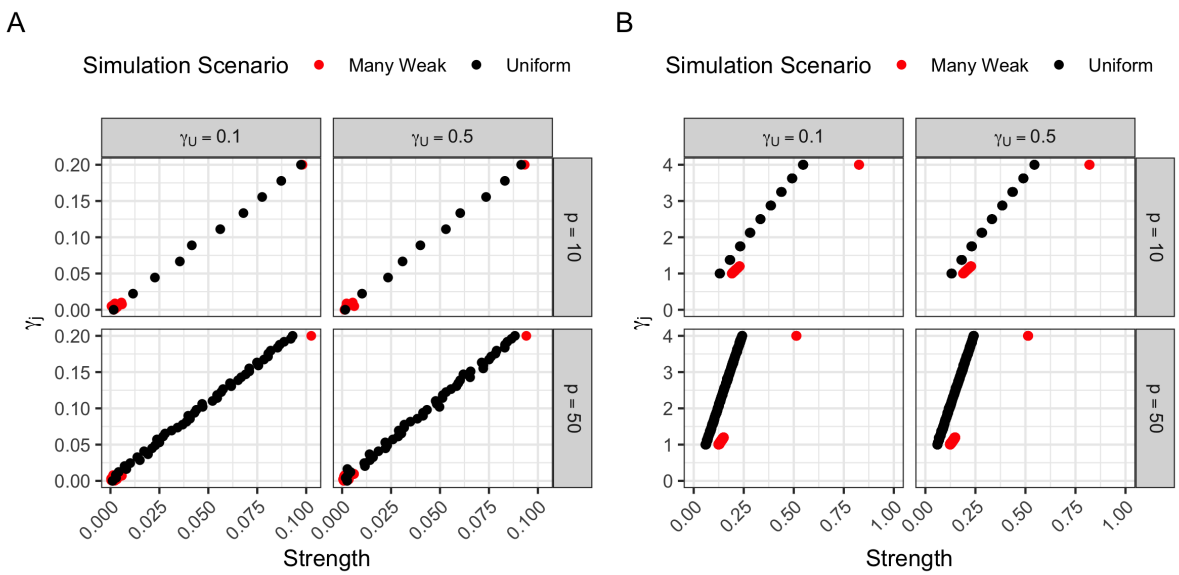

FIGURE 3 When p is larger, similar sized coefficients lead to lower strength. The effect is smaller when we are in a scenario where one coefficient is relatively much larger than the rest, rather than when the coefficients are evenly spread out. A: Scenarios 1 and 3. B: Scenarios 2 and 4.

We take a moment to explain the differences between our result in Section 3.1 with a single instrument with $\gamma_1 = 4$ and our results in this section where one of the instruments has $\gamma_i = 4$, but others have much smaller γ's. We see that if the variation in the exposure model is determined by multiple independent instruments, the effect of one single instrument on producing an informative bound greatly diminishes. Specifically, Figure 7 shows that in a setting where we would be able to detect the direction of the ATE from an instrument with $\gamma_i = 4$ if only $p = 10$ instruments are contributing to the exposure, that same coefficient would not be large enough if $p = 50$ instruments were contributing. This is a result of an altered relationship between γ_i and ST as seen on Figure 3; the strength of a single instrument is smaller when many are present. This suggests that for exposures that are determined by many instruments, the strongest among these instruments must be even stronger for a bound-based analysis to be useful. In other words, multiple instruments may not be helpful in a bound-based analysis when the exposure is polygenic in nature.

Our results also have dire implications when some instruments turn out to be invalid. If, as suggested by [22], we take the union of IV bounds so that the union bound is guaranteed to cover the true ATE so long as there is at least one valid instrument, the union bound will likely be non-informative because there was at least one IV bound in our scenario that was non-informative. Without making some assumptions about the nature of the invalid IVs, it would generally be infeasible to obtain useful information from a bound-based analysis.

Overall, combining our results in Section 3.1, our conclusion about using nonparametric IV bounds in two-sample MR studies is grim. Such a nonparametric analysis would require very strong instruments and/or effect sizes, which are rare in MR studies, and even stronger than those in one-sample settings. Also, multiple instruments are no better than having a single, strong instrument.

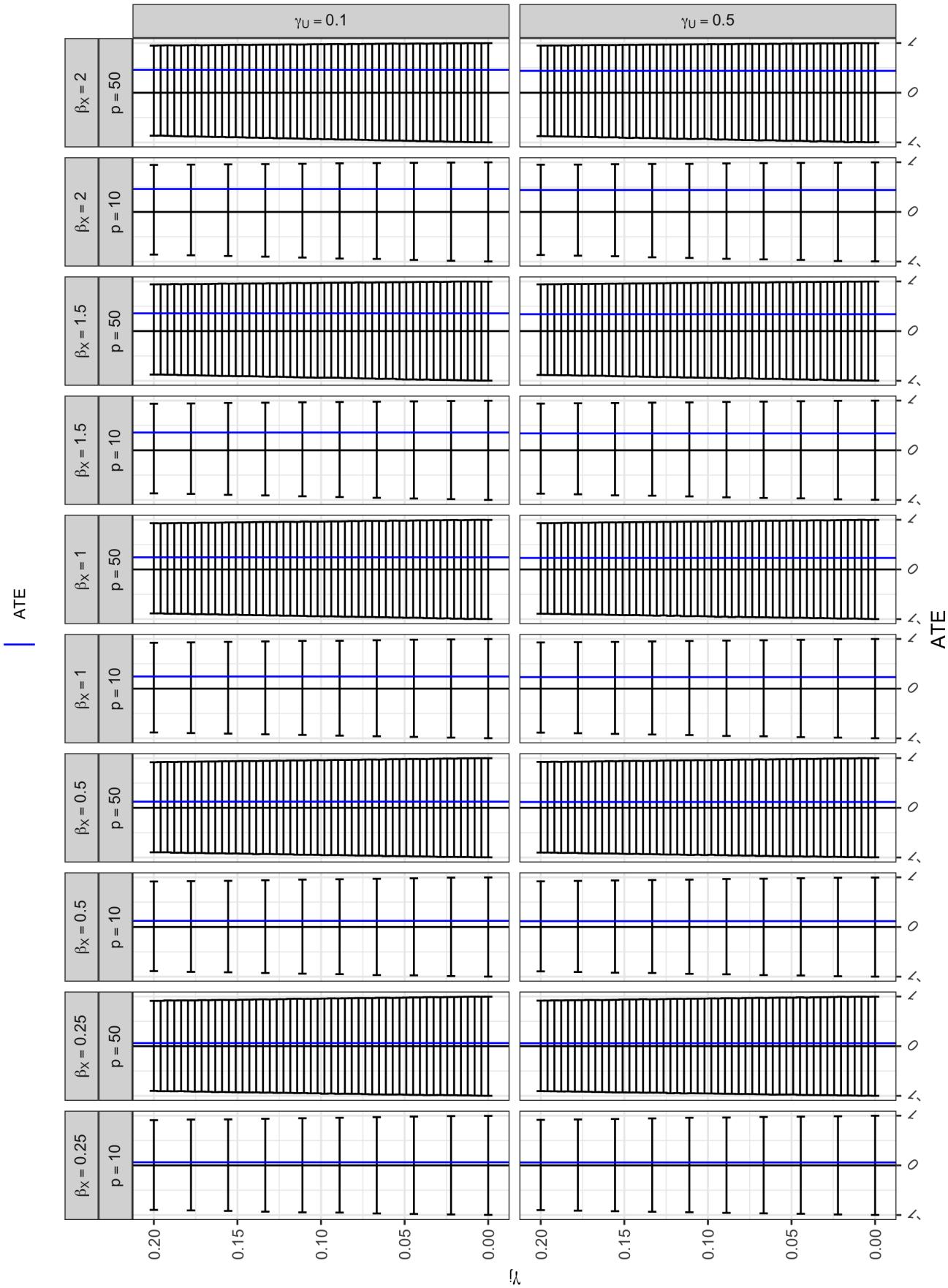

FIGURE 4 Bounds based on monte carlo integration with 1,000,000 resamples in scenario 1.

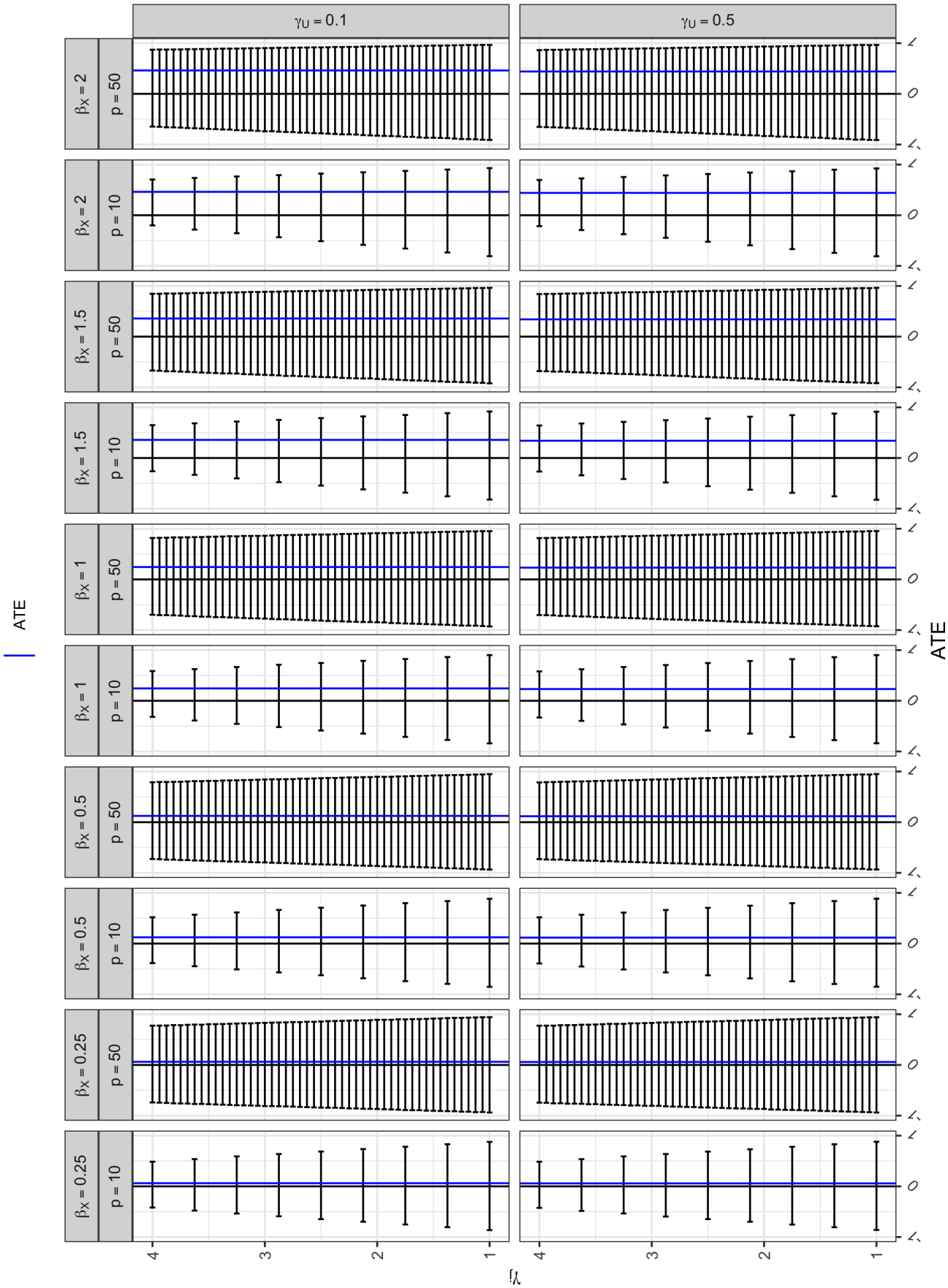

FIGURE 5 Bounds based on monte carlo integration with 1,000,000 resamples in scenario 2.

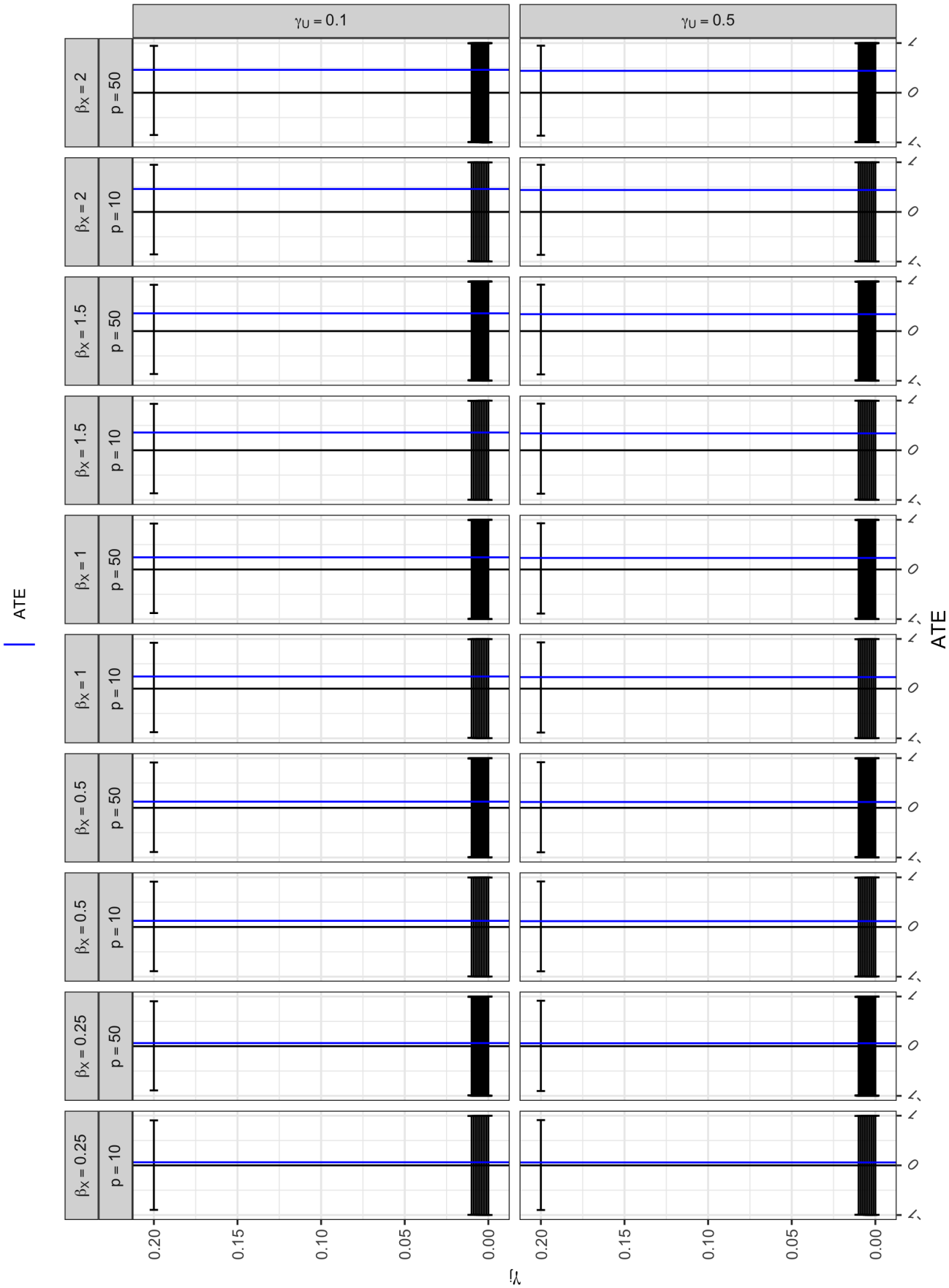

FIGURE 6 Bounds based on monte carlo integration with 1,000,000 resamples in scenario 3.

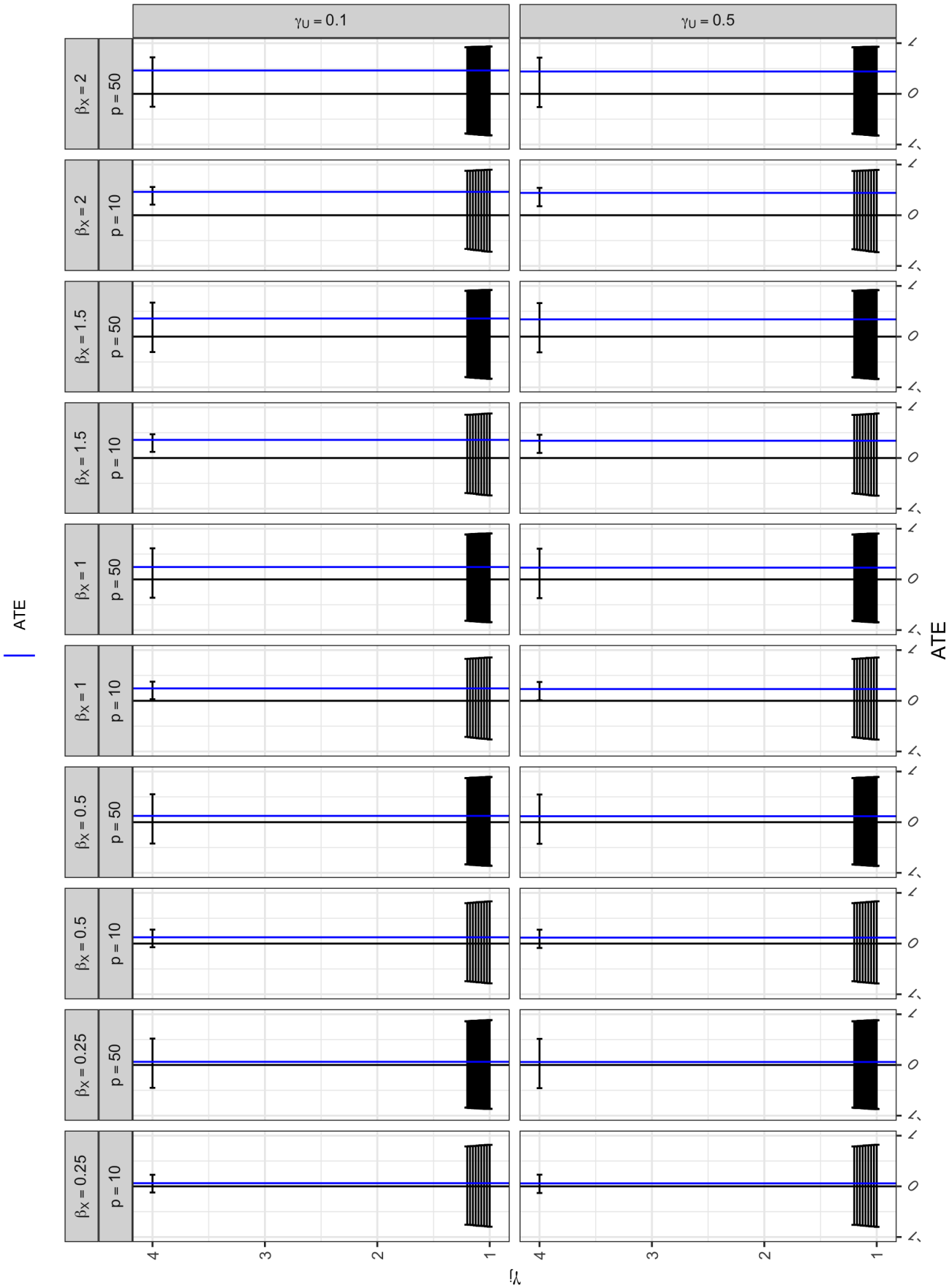

FIGURE 7 Bounds based on monte carlo integration with 1,000,000 resamples in scenario 4.

4 | CHARACTERING THE LOSS OF INFORMATION IN TWO-SAMPLE MR STUDIES

As hinted in Theorem 1, the increase in the bound's length is an inevitable "cost" of using two-sample designs instead of one-sample designs in MR studies. This section investigates this loss of information by creating a plausible range of the joint distribution of the outcome and the exposure given the instrument, $P(X = x, Y = y|Z = z)$, based on the observed data from two-sample MR studies. Specifically, using $P(X = x|Z = z)$, $P(Y = y|Z = z)$, and a uniform prior over the unknown quantity $\text{Cov}(X = x, Y = y|Z = z)$ that lead to satisfying the IV assumptions, we compute $P(X = x, Y = y|Z = z)$ and its corresponding one-sample IV bounds from[13] and[16]; see Appendix D for details on this procedure. If a large number of one-sample IV bounds obtained from this procedure do not cover zero, then there is some evidence for a non-zero exposure effect and a one-sample MR study may yield informative bounds on the ATE. However, if a large number of the one-sample IV bounds cover zero, there is little hope of obtaining information about the ATE from bound-based analyses even if we used a one-sample MR design; in other words, the one-sample IV bounds are likely to be equally conservative as the two-sample IV bounds.

Table 1 presents nine different sets of values of the marginal distributions $P(Y|Z)$ and $P(X|Z)$ that investigators could theoretically obtain from hypothetical two-sample MR studies. Figure 8 shows the one-sample IV bounds from the procedure we illustrated above.

Row A of Figure 8 shows three scenarios where the two-sample bounds are all centered close to zero with similar widths. But, the conclusions from the one-sample bound analysis are rather different. Column 1 shows no one-sample bounds would allow us to determine the presence of a non-zero exposure effect. Column 2 indicates that about 26.3% of the one-sample IV bounds do not contain 0 while for column 3 that number is approximately 35.9%. However, the latter includes one-sample bounds entirely above and below 0.

Row B illustrates three scenarios where the two-sample bounds are centered well above zero and have large widths. We see one case where we have no hope of determining direction of the ATE from the one-sample bounds (column 1), one case where we are most likely to determine the ATE's direction (column 2), and one case where we are unlikely to determine the ATE's direction (column 3).

Row C is similar to row A in that all the two-sample bounds are centered around 0, but the widths of the two-sample bounds are narrow. The three columns indicate similar conclusions as row A, showing that even with rather narrow two-sample bounds centered around 0, the one-sample bounds may still reveal some information about presence as well as the direction of the exposure effect.

Overall, the procedure and the examples above show that some two-sample MR studies could potentially reveal something useful about the ATE had we used a one-sample design. Nevertheless, we mention a word of caution when interpreting the results above, especially concerning the flat prior on the covariances. For example, a scenario like the one resulting in the bounds presented in row B, column 2 only provides honest information about the one-sample bounds if our prior on $\text{Cov}(X, Y|Z)$ is correctly specified. If the prior is mis-specified whereby most one-sample bounds cover negative values of the ATE, a negative value of the ATE is possible. But in this case, if the ATE is in fact negative, the procedure does rule out the possibility of one-sample bounds being able to ascertain this because all one-sample bounds covering a negative ATE also covers 0.

5 | USING BOUND-BASED ANALYSIS IN TWO, POSITIVE CONTROL EXAMPLES

We demonstrate our findings about the behavior of two-sample IV bounds on two real MR studies. Our first study examines the effect of smoking on incidence of lung cancer and our second study examines the effect of self-reported high cholesterol on incidence of heart attack. The effect of smoking on lung cancer is known to be strong and positive[45]. Also, while the exact mechanism between high cholesterol and heart disease is still being discussed[46,47], some meta-analyses of randomized clinical trials of the effect of cholesterol-lowering medication suggest a strong causal relationship[48,37]. In both cases, we assess what conclusions can be obtained by using bound-based analyses in studies where the causal effects are known to be strong and positive.

The study data were obtained from the UK Biobank data stored in the Integrative Epidemiology Unit (IEU) GWAS database. We use the `TwoSampleMR` R package[49] with the recommended defaults to extract and clean the data. For more details, see Appendix E.

For the effect of smoking on lung cancer, we used 84 genetic instruments, and for the effect of cholesterol on heart attack, we used 54 genetic instruments. The average instrument strengths were 0.0042 (range: 0.0032 to 0.0091) for smoking and 0.0005

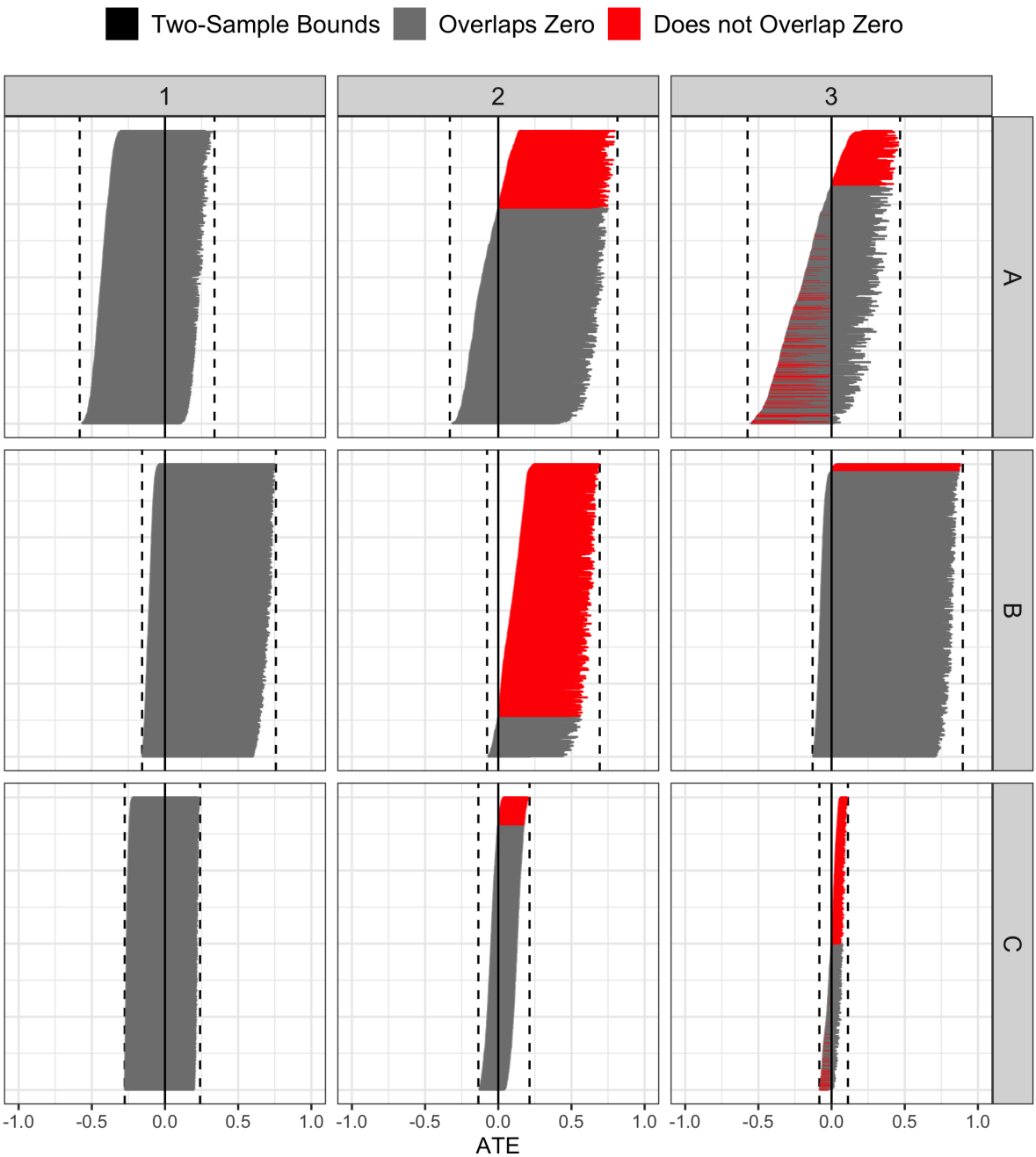

FIGURE 8 One-sample bounds (horizontal lines) and two-sample bounds (vertical dotted lines). Red color represents one-sample bounds that do not cover zero and gray color represents one-sample bounds that do cover zero.

(range: 0.0002 to 0.0022) for cholesterol; these values are much smaller than the ST = 0.5 needed to guarantee bounds with length less than 1. As such, the two-sample bounds in Figure 9 are wide; all of them have width greater than 1 and they convey no information about the causal effects of interest. Additionally, using our method from Section 4, the direction of the ATE may be difficult to determine had we used a one-sample design; see Figure 10. Appendix E contains additional analysis, notably demonstrating that aggregating multiple bounds through intersections are also non-informative.

Overall, while nonparametric bounds allow us to not make parametric assumptions frequent in two-sample MR analyses, they may provide little, if any, information about the exposure effects, even if the exposure effect is known to be positive and strong.

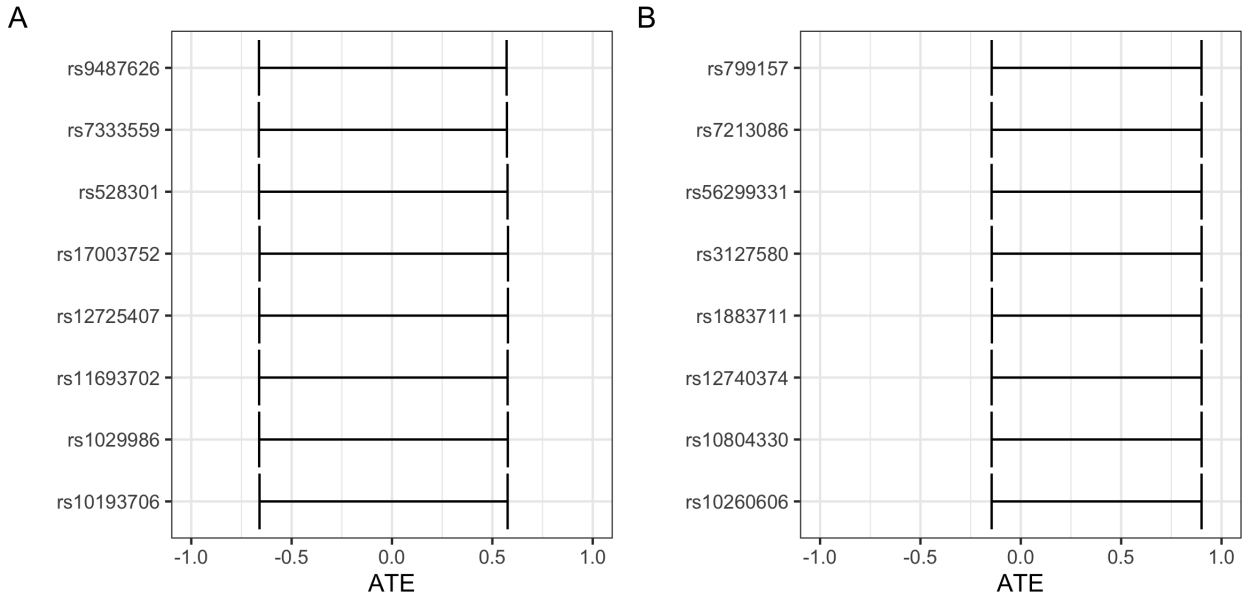

FIGURE 9 Two-sample IV bounds for the two real data examples with 8 SNPs from each data set. A: Two-sample IV bounds for the ATE of smoking on the incidence of lung cancer. B: Two-sample IV bounds for the ATE of high cholesterol on the incidence of heart attack.

Additionally, since many two-sample MR studies involve weak instruments, we believe bound-based approaches will likely have limited practical value to uncover causal effects.

6 | DISCUSSION

Nonparametric bounds are without a doubt an attractive concept. With a minimal set of assumptions, they let investigators obtain bounds on the average treatment effect. However, as we have seen above, in typical MR studies with two-sample summary data, a bound-based analysis may generally be uninformative for two reasons. First, while IV bounds in one-sample settings have length always less than 1, in two-sample settings, this is not always the case, and the bounds are often more conservative. Second, many genetic variants in MR studies are too weakly associated with the exposure to produce bounds with length less than 1 or bounds that exclude 0. Indeed, our two real data examples showed that despite having strong causal effects, bound-based analyses were unable to detect these effects.

We also outlined an approach to roughly quantify the information loss going from one-sample designs to two-sample designs and to assess the range of conclusions that can be drawn if we had one-sample data. We demonstrate our method to a few different settings of two-sample data and showed the range of conclusions that can be drawn about the ATE.

Overall, our general recommendation is that unless investigators have a very strong instrument, ideally exceeding ST > 0.5, bounds will unlikely be useful as a nonparametric analysis of the ATE, even with multiple instruments. Even if ST > 0.5, one would need strong IVs and/or strong effect sizes to make sure that the bounds do not cover 0. Finally, investigators can use our procedure above to assess whether it is worthwhile to use a one-sample MR design over a more typical (and arguably easier) two-sample MR design as the bounds under a one-sample design is generally less conservative than bounds from a two-sample design. Nevertheless, there may be few limited, but meaningful use cases for using bounds to study the ATE in two-sample MR studies; see[36] for one example. First when one has prior knowledge about the direction of the effect, but wish to get a better sense of the magnitude, nonparametric bounds can provide an upper limit on this magnitude. For example, when the exposure is known to cause harm or benefit, say in our smoking example, an upper bound on this effect would tell investigators about the maximum possible effect that smoking could have on increasing the incidence of lung cancer. Second, two-sample IV bounds

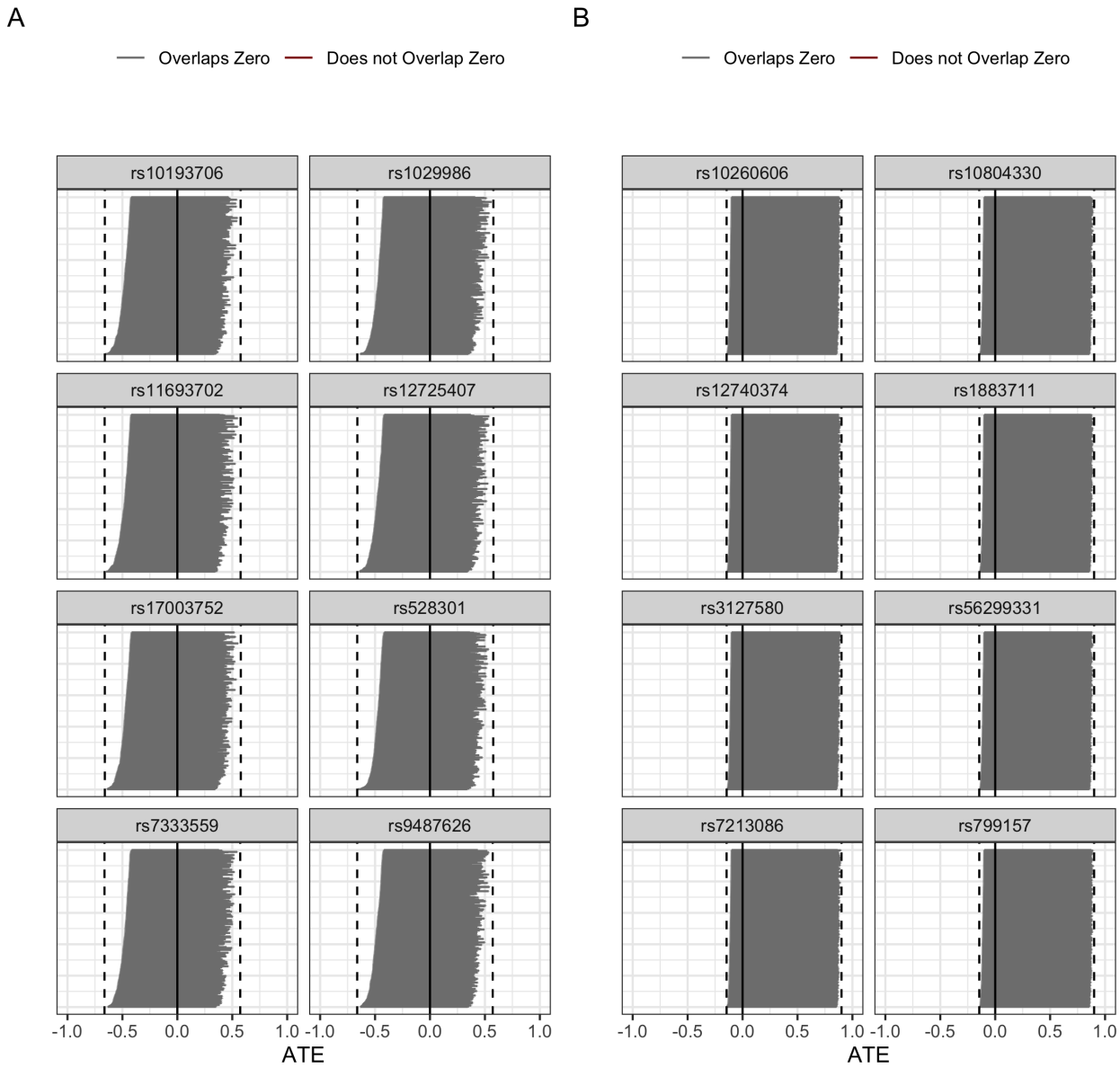

FIGURE 10 Potential one-sample IV bounds for the two real data examples using the method described in Section 4. A: One-sample IV bounds for the ATE of smoking on the incidence of lung cancer from 500 potential one-sample distributions for each SNP. B: One-sample IV bounds for the ATE of high cholesterol on the incidence of heart attack from 500 potential one-sample distributions for each SNP.

can be used to check estimates from parametric methods to see if they lie inside of the bounds; if the estimates lie outside of the bounds, then the parametric models underlying the estimates are likely mis-specified.

References

1. Davey Smith G, Ebrahim S. 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?. *International Journal of Epidemiology* 2003; 32(1): 1–22. doi: 10.1093/ije/dyg070

2. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. *Statistics in Medicine* 2008; 27(8): 1133–1163. doi: 10.1002/sim.3034

3. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Boca Raton: Chapman and Hall/CRC. 1st edition ed. 2015.

4. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis with Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* 2013; 37(7): 658–665. doi: 10.1002/gepi.21758

5. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, EPIC- InterAct Consortium . Using Published Data in Mendelian Randomization: A Blueprint for Efficient Identification of Causal Risk Factors. *European Journal of Epidemiology* 2015; 30(7): 543–552. doi: 10.1007/s10654-015-0011-z

6. Davies NM, Holmes MV, Smith GD. Reading Mendelian Randomisation Studies: A Guide, Glossary, and Checklist for Clinicians. *BMJ* 2018; 362. doi: 10.1136/bmj.k601

7. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the Suitability of Summary Data for Two-Sample Mendelian Randomization Analyses Using MR-Egger Regression: The Role of the I2 Statistic. *International Journal of Epidemiology* 2016; 45(6): 1961–1974. doi: 10.1093/ije/dyw220

8. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* 2016; 40(4): 304–314. doi: 10.1002/gepi.21965

9. Verbanck M, Chen CY, Neale B, Do R. Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nature Genetics* 2018; 50(5): 693–698. doi: 10.1038/s41588-018-0099-7

10. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score. *Annals of Statistics* 2020; 48(3): 1742–1769. doi: 10.1214/19-AOS1866

11. Burgess S, Small DS, Thompson SG. A Review of Instrumental Variable Estimators for Mendelian Randomization. *Statistical Methods in Medical Research* 2017; 26(5): 2333–2355. doi: 10.1177/0962280215597579

12. Slob EAW, Burgess S. A Comparison of Robust Mendelian Randomization Methods Using Summary Data. *Genetic Epidemiology* 2020; 44(4): 313–329. doi: 10.1002/gepi.22295

13. Balke A, Pearl J. Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association* 1997; 92(439): 1171–1176. doi: 10.1080/01621459.1997.10474074

14. Cheng J, Small DS. Bounds on Causal Effects in Three-Arm Trials with Non-Compliance. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 2006; 68(5): 815–836.

15. Manski CF. Nonparametric Bounds on Treatment Effects. *The American Economic Review* 1990; 80(2): 319–323.

16. Richardson TS, Robins JM. ACE Bounds; SEMs with Equilibrium Conditions. *Statistical Science* 2014; 29(3): 363–366. doi: 10.1214/14-STS485

17. Robins JM. The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using A New Approach to Causal Inference in Longitudinal Studies. *Health Service Research Methodology: A Focus On AIDS* 1989: 113–159.

18. Ramsahai RR. Causal Bounds and Observable Constraints for Non-Deterministic Models. *J. Mach. Learn. Res.* 2012; 13: 829–848.

19. Swanson SA, Hernán MA, Miller M, Robins JM, Richardson TS. Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes. *Journal of the American Statistical Association* 2018; 113(522): 933–947. doi: 10.1080/01621459.2018.1434530

20. Swanson SA, Hernán MA. Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome). *Epidemiology* 2013; 24(3): 370–374. doi: 10.1097/EDE.0b013e31828d0590

21. Didelez V, Sheehan N. Mendelian Randomization as an Instrumental Variable Approach to Causal Inference. *Statistical Methods in Medical Research* 2007; 16(4): 309–330. doi: 10.1177/0962280206077743

22. Swanson SA. Commentary: Can We See the Forest for the IVs? Mendelian Randomization Studies with Multiple Genetic Variants. *Epidemiology* 2017; 28(1): 43–46. doi: 10.1097/EDE.0000000000000558

23. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.. *Journal of Educational Psychology* 1974; 66(5): 688–701. doi: 10.1037/h0037350

24. Splawa-Neyman J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.. *Statistical Science* 1923; 5(4): 465–472. Translated in 1990.

25. Cox DR. *Planning of Experiments*. Planning of ExperimentsOxford, England: Wiley . 1958.

26. Rubin DB. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* 1980; 75(371): 591–593. doi: 10.2307/2287653

27. Wang L, Tchetgen Tchetgen E. Bounded, Efficient and Multiply Robust Estimation of Average Treatment Effects Using Instrumental Variables. *arXiv:1611.09925 [stat]* 2018.

28. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 1996; 91(434): 444–455. doi: 10.2307/2291629

29. Baiocchi M, Cheng J, Small DS. Instrumental Variable Methods for Causal Inference: Instrumental Variable Methods for Causal Inference. *Statistics in Medicine* 2014; 33(13): 2297–2340. doi: 10.1002/sim.6128

30. Stock JH, Wright JH, Yogo M. A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics* 2002; 20(4): 518–529. doi: 10.1198/073500102288618658

31. Burgess S. Sample Size and Power Calculations in Mendelian Randomization with a Single Instrumental Variable and a Binary Outcome. *International Journal of Epidemiology* 2014; 43(3): 922–929. doi: 10.1093/ije/dyu005

32. Verma A, Bradford Y, Dudek S, et al. A Simulation Study Investigating Power Estimates in Phenome-Wide Association Studies. *BMC Bioinformatics* 2018; 19(1): 120. doi: 10.1186/s12859-018-2135-0

33. Millard LAC, Munafò MR, Tilling K, Wootton RE, Smith GD. MR-pheWAS with Stratification and Interaction: Searching for the Causal Effects of Smoking Heaviness Identified an Effect on Facial Aging. *PLOS Genetics* 2019; 15(10): e1008353. doi: 10.1371/journal.pgen.1008353

34. King C, Mulugeta A, Nabi F, Walton R, Zhou A, Hyppönen E. Mendelian Randomization Case-Control PheWAS in UK Biobank Shows Evidence of Causality for Smoking Intensity in 28 Distinct Clinical Conditions. *EClinicalMedicine* 2020; 26. doi: 10.1016/j.eclinm.2020.100488

35. Burgess S, Thompson SG. Improving Bias and Coverage in Instrumental Variable Analysis with Weak Instruments for Continuous and Binary Outcomes. *Statistics in Medicine* 2012; 31(15): 1582–1600. doi: 10.1002/sim.4498

36. Diemer EW, Labrecque J, Tiemeier H, Swanson SA. Application of the Instrumental Inequalities to a Mendelian Randomization Study With Multiple Proposed Instruments. *Epidemiology* 2020; 31(1): 65. doi: 10.1097/EDE.0000000000001126

37. Cholesterol Treatment Trialists' (CTT) Collaborators . The Effects of Lowering LDL Cholesterol with Statin Therapy in People at Low Risk of Vascular Disease: Meta-Analysis of Individual Data from 27 Randomised Trials. *The Lancet* 2012; 380(9841): 581–590. doi: 10.1016/S0140-6736(12)60367-5

38. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *JNCI: Journal of the National Cancer Institute* 1959; 22(1): 173–203. doi: 10.1093/jnci/22.1.173

39. Yusuf S, Hawken S, Ounpuu S, et al. Obesity and the Risk of Myocardial Infarction in 27,000 Participants from 52 Countries: A Case-Control Study. *Lancet (London, England)* 2005; 366(9497): 1640–1649. doi: 10.1016/S0140-6736(05)67663-5

40. Schmitt J, Seidler A, Diepgen TL, Bauer A. Occupational Ultraviolet Light Exposure Increases the Risk for the Development of Cutaneous Squamous Cell Carcinoma: A Systematic Review and Meta-Analysis. *Br J Dermatol* 2011; 164(2): 291–307. doi: 10.1111/j.1365-2133.2010.10118.x

41. Loh PR, Bhatia G, Gusev A, et al. Contrasting Genetic Architectures of Schizophrenia and Other Complex Diseases Using Fast Variance-Components Analysis. *Nature Genetics* 2015; 47(12): 1385–1392. doi: 10.1038/ng.3431

42. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* 2016; 99(1): 139–153. doi: 10.1016/j.ajhg.2016.05.013

43. Nj T, Cmt G, N S, Dj L, Jb R. Genetic Architecture: The Shape of the Genetic Contribution to Human Traits and Disease.. *Nature reviews. Genetics* 2017; 19(2): 110–124. doi: 10.1038/nrg.2017.101

44. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs Explain a Large Proportion of the Heritability for Human Height. *Nature Genetics* 2010; 42(7): 565–569. doi: 10.1038/ng.608

45. on Smoking SGAC, ealth, United States. . *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Public Health Service PublicationU.S. Department of Health, Education, and Welfare, Public Health Service . 1964.

46. Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian Randomization of Blood Lipids for Coronary Heart Disease. *European Heart Journal* 2015; 36(9): 539–550. doi: 10.1093/eurheartj/eht571

47. Richardson TG, Sanderson E, Palmer TM, et al. Evaluating the Relationship between Circulating Lipoprotein Lipids and Apolipoproteins with Risk of Coronary Heart Disease: A Multivariable Mendelian Randomisation Analysis. *PLOS Medicine* 2020; 17(3): e1003062. doi: 10.1371/journal.pmed.1003062

48. Cholesterol Treatment Trialists' (CTT) Collaborators . Efficacy and Safety of Cholesterol-Lowering Treatment: Prospective Meta-Analysis of Data from 90 056 Participants in 14 Randomised Trials of Statins. *The Lancet* 2005; 366(9493): 1267–1278. doi: 10.1016/S0140-6736(05)67394-1

49. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 2018; 7: e34408. doi: 10.7554/eLife.34408

TABLE 1 Values of $P(X = 1|Z = z)$ and $P(Y = 1|Z = z)$ used to illustrate our approach. For each cell (e.g. row A, column 1), we have $\{P(X = 1|Z = 0), P(X = 1|Z = 1), P(X = 1|Z = 2)\}$ on the first row and $\{P(Y = 1|Z = 0), P(Y = 1|Z = 1), P(Y = 1|Z = 2)\}$ on the second row.

	Column 1	Column 2	Column 3
Row A	{0.125, 0.399, 0.080}	{0.244, 0.275, 0.185}	{0.603, 0.469, 0.310}
	{0.699, 0.840, 0.742}	{0.238, 0.089, 0.146}	{0.638, 0.346, 0.719}
Row B	{0.886, 0.968, 0.874}	{0.139, 0.441, 0.334}	{0.901, 0.909, 0.935}
	{0.805, 0.822, 0.951}	{0.179, 0.359, 0.559}	{0.821, 0.810, 0.905}
Row C	{0.175, 0.079, 0.365}	{0.493, 0.911, 0.085}	{0.434, 0.045, 0.733}
	{0.599, 0.358, 0.087}	{0.360, 0.480, 0.441}	{0.747, 0.370, 0.169}