MINI-SENTINEL METHODS DEVELOPMENT

SIGNAL EVALUATION WORKING GROUP REPORT

Prepared by: *Section I:* Jeremy A. Rassen, ScD[1] and Jessica Myers, PhD[1] on behalf of the Signal Evaluation Working Group; *Section II:* Jeremy A. Rassen, ScD[1] and Sebastian Schneeweiss, MD, ScD[1]; *Section III:* Jessica A. Myers, PhD,[1] Jeremy A. Rassen, ScD,[1] Joshua J. Gagne, MS, PharmD,[1] Sebastian Schneeweiss, MD, ScD,[1] Krista F. Huybrechts, MS, PhD,[1] Kenneth J. Rothman, MPH, DrPH,[2] Marshall M. Joffe, MD, MPH, PhD,[3] Robert Glynn, ScD, PhD[1]

Author Affiliations: 1. Brigham and Women's Hospital and Harvard Medical School, Boston, MA 2. RTI International, Research Triangle Park, NC 3. University of Pennsylvania School of Medicine, Philadelphia, PA

August 12, 2011

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Methods Development

Signal Evaluation Working Group Report

I. INTRODUCTION

A. BACKGROUND

Signal evaluation in drug safety surveillance with non-randomized data requires strong confounding adjustment, since confounding by indication is likely to be substantial. In a distributed safety monitoring system like that envisioned for FDA's Sentinel Initiative, a robust and automated approach to confounding adjustment is crucial. Any automation must be built upon sound design principles and careful quality control but also be built for speed and scale.

Confounding control, whether through stratification, restriction, matching, or regression modeling, begins with identification of potential confounders and correct selection of the covariates that influence the medication use and outcome under evaluation. Traditional approaches—such as formal processes like drawing a directed acyclic graph (DAG)[1]—will likely not scale to the number of monitoring projects envisioned for the Sentinel System. Further, a standing library of covariate definitions, even one based on a common data model, may not include all the important risk factors for the study at hand or the same set of risk factors for every member of the distributed system, and may therefore require all participating entities to subscribe to a "lowest common denominator" of available data elements. The fact that drug user populations shift over time due to expansion of indications on one hand, or risk management plans on the other, will make it even more cumbersome to pick a single set of "correct" variables that are applicable longitudinally.

The ideal automated procedure for covariate selection would create and select pre-exposure covariates and achieve confounding control as well as or better than a team of investigators could. This process would enable valid study of treatment/outcome associations with a minimum of investigator attention or intervention,[2] making monitoring fast and scalable. In this report, we investigate whether one procedure, the high-dimensional propensity score (hd-PS) algorithm, could serve as an automated mechanism for confounding adjustment for a medical product active safety surveillance system, and describe the strengths and limitations of that proposal. We discuss the robustness of the covariate selection process, the computation required, potential for over- or under-selection of variables, practical considerations in a distributed database network, and cases where automated confounding adjustment may not function optimally.

Any automated procedure also requires not adjusting for certain types of variables that have been demonstrated to increase rather than reduce bias. In particular, recent literature has questioned whether instrumental variables (IVs or instruments) or instrument-like variables should be adjusted for. IVs are variables that are associated with exposure but are not associated with outcome except through their effect on exposure. Although IVs may be used to provide an unbiased estimate of exposure effect in the presence of unmeasured confounding, theoretical and simulation work has found that conditioning on an IV in effect estimation may increase bias. However, the implication of these results for automated confounder selection remains unclear. True instruments are difficult to identify and their assumptions are not empirically verifiable.[3] Further, the available theoretical studies provide results within a narrow range of generally extreme cases. Since any increase in bias due to unnecessary conditioning must be weighed against the danger of excluding real confounders from the conditioning set, we set out to investigate the level to which IVs, or IV-like variables, would affect the validity of signal evaluation in a medical product active safety surveillance system across a range of possible circumstances.

To that end, we focus on the case where an IV may exist in the set of measured variables but is uncertain to investigators. We consider both true instruments—variables with no direct effect on unobserved confounding factors or outcome—and 'near-instruments'—variables that are weakly associated with the unmeasured confounding. We also explore effects under varying assumptions about the strength of the IV association with exposure and the magnitude of the unmeasured confounding.

B. OVERVIEW & CONCLUSIONS OF REPORT

The report is presented in two sections. The first section provides a rationale for the use of automated covariate adjustment algorithms in claims database studies and reports on experiences with hd-PS as it has been applied in studies to date. The second section investigates Z-bias in detail.

We believe this report demonstrates that the high-dimensional propensity score (hd-PS) can serve as a valuable tool in a distributed medical product safety surveillance system, even at the expense of selecting variables that introduce Z-bias. Z-bias, where we observed it, was minimal when compared to bias due to unmeasured confounding. In non-randomized pharmacoepidemiology using healthcare databases, confounding bias is generally considered to be the greatest threat to study validity. An automated confounding adjustment system that generates and prioritizes a large number of covariates, even with somewhat imperfect variable selection, should yield a collection of variables that, if adjusted for, will improve study validity far more than it will harm it.

Our findings are encouraging with respect to the substantial challenges in developing strategies for automated covariate adjustment needed in a large-scale distributed medical product safety surveillance system. The strategies that we outline should allow for valid, robust, and scalable confounding adjustment across numerous simultaneous investigations.

II. AUTOMATED COVARIATE ADJUSTMENT IN A DISTRIBUTED MEDICAL PRODUCT SAFETY SURVEILLANCE SYSTEM

A. INTRODUCTION

Distributed safety monitoring systems such as that envisioned for FDA's Sentinel Initiative will benefit from automating large parts of today's pharmacoepidemiology study process. They will require automation that is built upon sound design principles and careful quality control, and also built for speed and scale. Speed is required for fast evaluation and identification of safety signals, while scale is required to manage both the number of patients under observation and the number of potential safety signals the systems need to be able to evaluate. In this setting, the intelligence that investigators normally apply study-by-study needs to be encapsulated in study frameworks and algorithms that can be applied reliably and in a largely hands-off fashion. We discuss automating confounding adjustment in longitudinal healthcare databases, a likely data source for active safety monitoring activities.

Adequate and appropriate study design provides a certain portion of the necessary intelligence. Self-controlled designs, which minimize confounding by comparing a patient to him or herself, are well-suited to studies of acute onset events such as allergic reactions or situations of time-varying exposures.[4] However, in the majority of monitoring scenarios, cohort designs and their related sampling strategies, like nested case-control or case-cohort designs, will be better suited. We have suggested that a successful drug safety cohort study will employ an incident user design[5] with well-defined covariate assessment and exposure definition windows, matched analyses, and other familiar components.[6] An important part of cohort-type designs is choosing a suitable comparison group, often an active comparator. This choice is critical to study validity and will need to be decided by study personnel on a case-by-case basis.

Any cohort study requires careful control for between-person confounding. Confounding control, whether through stratification, restriction, matching, or regression modeling, begins with identification of potential confounders and correct selection of the covariates that influence the medication use and outcome under evaluation. Traditionally, this is done by applying subject matter expertise; it may be augmented by a more formal process such as drawing a directed acyclic graph (DAG).[1] Covariates may be created (identified and coded) specifically for the study or, in pharmacoepidemiology studies using healthcare databases, covariate definitions may already exist in a standing library. However, this traditional approach does not scale well to either the large number of covariates that is increasingly seen in pharmacoepidemiologic studies nor to the number of monitoring projects envisioned for an active medical product surveillance system. A standing library of covariate definitions, even one based on a common data model, may not include all the important risk factors for the study at hand and may require all participating entities to subscribe to a "lowest common denominator" of available data elements. The fact that drug user populations shift over time due to expansion of indications on one hand, or risk management plans on the other, will make it even more cumbersome to pick a single set of "correct" variables that are applicable on an ongoing basis.

The ideal automated procedure for covariate selection would create and select pre-exposure covariates and achieve confounding control as well as or better than a team of investigators could. This process would enable valid study of treatment/outcome associations with a minimum of investigator attention or intervention,[2] making monitoring fast and scalable. In this report, we investigate whether one

procedure, the hd-PS algorithm, could serve as an automated mechanism for confounding adjustment for a medical product active safety surveillance system, and describe the strengths and limitations of that proposal. In particular, we discuss the robustness of the covariate selection process, the computation required, potential for over- or under-selection of variables, practical considerations in a distributed database network, and cases where automated confounding adjustment may not function optimally.

B. THE HD-PS ALGORITHM

In earlier work, we described the high-dimensional propensity score (hd-PS), an automated covariate creation, selection, and adjustment process. It has now been applied to a number of pharmacoepidemiology studies.[7-14] **Table 1** shows what we consider to be evidence of the method's success: across a range of studies, a largely monotonic progression of the point estimate as additional levels of confounding control are applied. We observed that this progression may move towards a null finding, away from the null, or even to and beyond the null depending on the nature of the residual confounding. For example, in the first row, the unadjusted point estimate indicated that Cox-2 inhibitors (coxibs) are associated with a 9% increase of incidence of GI bleed as compared to non-selective non-steroidal anti-inflammatory drugs (ns-NSAIDs). Randomized trials suggested a lowering of risk by about 20% among healthy patients,[15,16] so we believe the unadjusted value to be upwardly biased. Adjusting for age, sex, and other basic variables moves the point estimate downward toward the estimate predicted by the trial, with a 1% increase in risk. Further adjustment by other investigator-specified variables moves the estimate downward, to a 6% relative risk reduction. Applying hd-PS moves the estimate downward and yields the estimate closest to the expected value in the trial, to a 12 to 13% relative risk reduction in our routine care population.

Since the publication of the original algorithm, we have studied and made modifications to the procedure both to handle small study sizes[11] and to greatly improve speed.[17]

Table 1. hd-PS as applied to various studies. [Coxib: Cox-2 inhibitor; ns-NSAID: non-selective non-steroidal anti-inflammatory drug; TCA: tricyclic antidepressant; SSRI: selective serotonin uptake inhibitor; PPI: proton pump inhibitor; APM: anti-psychotic medication; MI: myocardial infarction; CV: cardiovascular]

Type of data source	Exposure (Referent)	Outcome	Relative Risk Estimate (95% Confidence Interval)				
			Unadjusted	Adjusted by Basic Variables	… Plus Other Pre-Specified Variables	… Plus hd-PS	Basic Variables and hd-PS Only
US Medicare claims data	Coxibs (ns-NSAIDS) [6]	GI bleed within 180 days	1.09 (0.91-1.30)	1.01 (0.84-1.21)	0.94 (0.78-1.12)	0.88 (0.73-1.06)	0.87 (0.72-1.05)
German claims data	Coxibs (ns-NSAIDS) [38]	GI bleed	1.21 (0.91-1.61)	-	0.99 (0.74-1.33)	0.67 (0.45-0.97)	-
UK claims data supplemented with EMR data	Coxibs (ns-NSAIDS)[12]*	Upper GI bleed	1.50	0.84	0.81	0.78	0.81
US Medicare claims data	Statins (glaucoma drugs)[6]	All-cause mortality within 180 days	0.56 (0.51-0.62)	0.77 (.069-0.85)	0.80 (0.70-0.90)	0.86 (0.76-0.98)	0.89 (0.78-1.02)
British Columbia claims data	TCAs (SSRIs), <18 y.o.[8]	Suicide within 1 year	0.59 (0.28–1.27)	0.66 (0.31-1.42)	0.71 (0.33–1.52)	0.92 (0.43-2.00)	-
British Columbia claims data	TCA (SSRIs), 18+ y.o.[9]	Suicide within 1 year	0.97 (0.77-1.21)	1.04 (0.83-1.31)	1.04 (0.82-1.31)	1.14 (0.88-1.47)	-
Mix of US Medicare, US commercial, and Canadian claims data	Clopidogrel + PPI (Clopidogrel alone) [39]	MI or CV death	1.74 (1.44-2.10)	1.62 (1.34-1.96)	1.32 (1.08-1.61)	1.22 (0.99-1.51)	-
US commercial claims data	Neurontin (Topamax)[10]	Suicide or attempted suicide	0.95 (0.76-1.19)	1.48 (1.17-1.87)	1.42 (1.11-1.80)	1.99 (1.45-2.73)	-
British Columbia claims data	Conventional APMs (Atypical APMs)[11]	Death	1.37 (1.11-1.69)	1.47 (1.13-1.90)	1.47 (1.14-1.91)	1.52 (1.14-2.02)	-
British Columbia claims data	Benzodiazepines (Atypical APMs)[11]	Death	1.37 (1.14-1.64)	1.52 (1.25-1.85)	1.28 (1.04-1.58)	1.20 (0.96-1.50)	-

C. APPLYING THE HD-PS ALGORITHM IN A MEDICAL PRODUCT ACTIVE SURVEILLANCE SYSTEM

The hd-PS algorithm creates and prioritizes covariates that may be confounders of the medical product-outcome association under study.[7] In its most common configuration, it takes as input the recorded history of medical encounters—the presence of diagnostic codes, procedure codes, hospitalizations, medication fills—experienced by the patient prior to exposure. The algorithm creates covariates from each of these binary events and assesses these new covariates for their association with exposure and with outcome. Using the Bross formula for confounding bias for binary variables,[19] it then ranks the group of covariates for their potential to bias the treatment/outcome under study and by default selects the top 500 of the covariates most likely to add bias to the study. It enters these variables into an exposure propensity score model and, after estimating the propensity score, the hd-PS algorithm initiates a fixed-ratio matching process that creates a cohort in which patients treated with the exposure and referent drugs are balanced with respect to measured covariates.

When using hd-PS, we recommend a set of basic design principles. In order to ensure clear temporality and to avoid other biases, we apply an incident user cohort design[6] in which exposure is required to be preceded by a term of non-use of the study drugs, thereby excluding prevalent users. In most cases, exposure should be contrasted with an active comparator group with the same indication; for example, users of coxibs would be compared to users of ns-NSAIDs rather than to non-users. All covariates that are assessed must be recorded within a defined period before the exposure date, often 180 or 365 days. Outcome is assessed during a limited follow-up time with censoring at the time of discontinuation or switching, in an as-treated analysis. These design criteria, while not the only ways to conduct a successful study, ensure that key epidemiologic principles are met: covariates are measured prior to exposure, incident users are compared "apples to apples" to other incident users, and exposure misclassification during follow-up is limited.

With this design in place, hd-PS can evaluate the data, identify and make a selection of covariates, estimate a propensity score, and match patients within the cohort, all without user intervention. Applied on a periodic basis over time, the result is a series of cohorts matched within each study location and matched on the best available covariate information at the time. As exposure or outcome frequency grows or the composition of the population using a drug changes over time, so will the covariates selected to optimally address confounding. While this is contrary to the principle of choosing covariates based on knowledge of the biologic processes at work, it is an effective approach in secondary data like the data used in a medical product active safety surveillance system. It is also pragmatic because it addresses the issue of change in drug usage patterns over time and thus ensures that maximum validity is achieved at each point in time and in each data environment, even across the heterogeneous data elements available in a distributed system.

Matching the cohorts imposes an implicit and useful restriction criterion: exposed patients are unmatched and thus excluded if no exchangeable referent patient exists and vice-versa. While analytically the results are similar to a trimmed propensity score approach,[20] a fixed-ratio matching process also provides other beneficial side-effects, such as a cohort that is balanced on all measured confounders and thus does not require further covariate adjustment in the outcome analysis. An inspection of the cohort's "Table 1" stratified by exposure category and data environment indicates, either visually or through the automated application of a measure like the Mahalanobis distance,[21,22]

residual imbalances that need to be addressed. The benefits of the simplicity of this balance quality check should not to be underestimated in a system that aspires to automate as many aspects as possible but which still allows for rapid quality checks of key elements of the underlying epidemiology.

D. DISCUSSION OF THE HD-PS'S STRENGTHS AND LIMITATIONS

Any automated process requires careful scrutiny of the process' strengths and limitations. For decades, epidemiologists have been taught that each confounding variable's relationship with the exposure and outcome must be fully understood on biological and medical-sociological grounds. Not surprisingly, a healthy skepticism is a common first reaction to an automated confounding adjustment approach. But context is important. In the case of a safety surveillance system based on secondary data, our biggest concern is unmeasured confounders because we are not in control of the data collection process and thus are prevented from defining necessary confounding variables from first principles.

In this section, we present some criticisms of hd-PS that have been voiced in the two years since the original publication of the algorithm and outline responses. In the end, we conclude that, despite minor limitations, hd-PS remains a valuable and reliable tool in most medical product safety surveillance activities.

1. General Issues with Variable Selection

An automated variable selection technique can fail in several ways: it can select too few confounders, too many covariates, and/or the wrong covariates.

With respect to the optimal number of covariates to be selected, we performed extensive simulation studies in which we sought to determine how many variables are required for maximal confounding adjustment achievable in a specific data source for a specific exposure-outcome pair. We examined hd-PS in both common and small study circumstances and determined that 350-400 variables are generally sufficient.[11] Since the algorithm selects the variables likely to cause the most bias first, we observed that point estimates ceased to change after 350-400 variables had been included in the propensity score. Selecting additional variables would likely do no harm, but selecting fewer is likely to lead to under-adjustment. Indeed, in a propensity score, including too many variables causes little harm as long as those variables are either confounders, proxies for confounders, or predictors of the outcome.[23] While a common concern from regression modeling is "overfitting" the propensity score model, a propensity score is an exposure prediction based on the data at hand; parameters in the propensity score model will not be interpreted or generalized to other datasets. Any co-linearity or so-called overfitting is not of relevance in this instance.[24,25]

Nevertheless, selection of the correct variables is important. While the Bross formula provides a reasonable way to quantify the potential bias if we fail to adjust for a particular variable, it does not consider more complex situations. In particular, certain variables such as colliders[26,27] may appear analytically to be confounders, but adjusting for them can increase rather than decrease bias. This increase comes from introducing so-called "backdoor paths" by which the influence of variables that are structurally unrelated to the treatment/outcome relationship is included in the treatment/outcome measure of association.[1]

Two relevant kinds of collider bias have been noted in the literature: M-bias, named for the shape of the DAG that characterizes it,[27] and Z-bias, named because the bias comes from the inclusion of an instrumental variable (often notated Z) in the analysis and also called "residual confounding amplification."[28,29] Types of collider bias that involve exposure and confounders[26] are not problematic in the case of an incident user cohort study because measuring confounders prior to exposure avoids possible collider constructs.

2. Variable Selection Issue: Potential for M-bias

M-bias (**Figure 1**) occurs from conditioning on an apparent confounder (C) which is actually a collider. C must be associated with two types of unmeasured confounders: a U1 that is associated only with the exposure and a U2 that is associated only with the outcome. Neither U1 nor U2 confound the exposure (X) to outcome (Y) association, so they should not be adjusted for. Evaluating C with the Bross formula may indicate potential for bias, because C will be associated with both X and Y via X's and Y's associations with U1 and U2, respectively. Adjusting for C will introduce bias due to a backdoor path from X to Y via the unmeasured confounders; this is a path that would have otherwise been closed had C not been adjusted for.

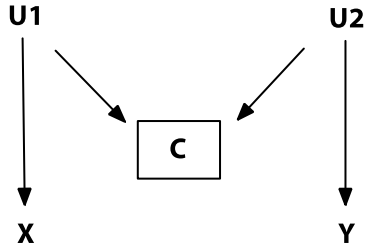


Figure 1. Example of M-bias.

Out of a concern that hd-PS's selection of colliders with this form could yield biased effect estimates, our colleagues undertook a simulation study to quantify the magnitude of that bias as it would manifest in common pharmacoepidemiology scenarios. Their findings indicate that the bias, while theoretically present, was generally small (<5%).[30] The authors pointed out that finding a scenario like this was rather difficult since any collider C would have to be a downstream consequence of both U1 and U2, which in an incident user design would also have to occur prior to exposure. Further, the strength of M-bias is limited to the strength of the weakest association in the chain of X→U1→C→U2→Y. At least one of these is likely to be weak because the hd-PS algorithm should capture all strong and moderately strong predictors of X and Y and thus diminish the strength of the residual confounding (U1 and U2) pathways. Our colleagues concluded, and we concur, that the bias, if present, would have a minimal impact in any realistic situation.

Moreover, outside the tidiness of a simulation environment, in other circumstances (**Figure 2**) the variable C may be a collider on one path (U1→C→U2), but a proxy for a confounder on another (U3). In this case, a DAG does not shed light on whether to adjust for C: the answer comes down to whether adjusting for C will do more good than harm. We believe that in most pharmacoepidemiology situations, we are unlikely to see true M-bias of any relative magnitude and, if we do, the complexity of the underlying biological and medical-sociological structures will likely yield a complicated situation that

mixes confounding and colliding. In the majority of these circumstances, we believe the reduction in bias from adjusting for the confounding will far offset any increase in bias due to conditioning on a collider.

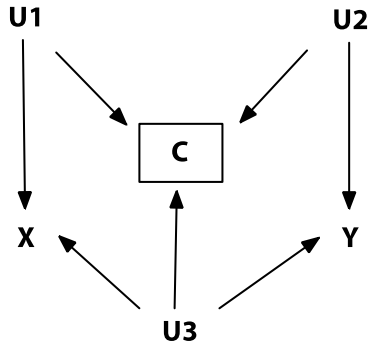


Figure 2. Example of M-bias plus confounding bias.

3. Variable Selection Issue: Potential for Z-bias

Z-bias refers to the bias caused by adjusting for an instrumental variable in studies that also have unmeasured confounding. An instrument is a variable that is associated only with the exposure and not with the outcome (other than through a pathway via exposure); such an instrument can serve to avoid bias if handled with the proper analytic tools.[31-33] It is not apparent from a DAG that adjusting for an instrument will cause harm but it has been shown theoretically[28] and empirically[29] that doing so will amplify existing residual confounding.

Our research group undertook a simulation study to quantify the effects of Z-bias; in particular, we sought to determine whether Z-bias was likely to be problematic in common pharmacoepidemiology circumstances.[34] Through a simulation analysis, Myers, et al. tested thousands of cases where Z-bias was expected to appear in varying degrees. In all cases, we found that Z-bias, while measureable, was only of substantial magnitude in cases of very strong unmeasured confounding. Further, even in these cases, the strongest Z-bias we observed represented less than 5% of the total study bias. This is demonstrated in **Figure 3**. The few instances with substantial observed Z-bias, represented by points above the diagonal line, are in cases where the bias due to unmeasured confounding ("crude bias") is strong. From this simulation analysis, we concluded that Z-bias was indeed a measureable phenomenon but that it was small in degree compared to studies' true threat to validity: unmeasured confounding. In cases where studies were affected by unmeasured confounding to lesser degrees, Z-bias was almost nonexistent.

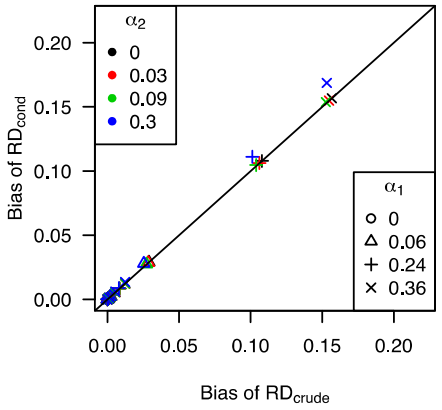


Figure 3. Observed Z-bias in simulation studies. Points along the diagonal line show no Z-bias, whereas points above the line indicate introduction of Z-bias by adjusting for an instrument. Figure republished from Myers, et al.[34]

In the end, since we will never be able to distinguish a confounder from a collider based on inspection of data, some effects of M- and Z-bias are unavoidable. Simulations, practice, and pragmatism tell us that adjusting for a variable that appears to be a confounder will in most circumstances decrease bias. In non-randomized pharmacoepidemiology, confounding bias is generally considered to be the greatest threat to study validity; any confounding bias will likely be of greater magnitude than collider bias and, on a covariate-by-covariate basis, confounders are likely to far outnumber colliders. Taken holistically, an automated confounding adjustment system that selects a large number of covariates, even with somewhat imperfect variable selection, should yield a collection of variables that, if adjusted for, will improve study validity far more than harm it (**Figure 4**).

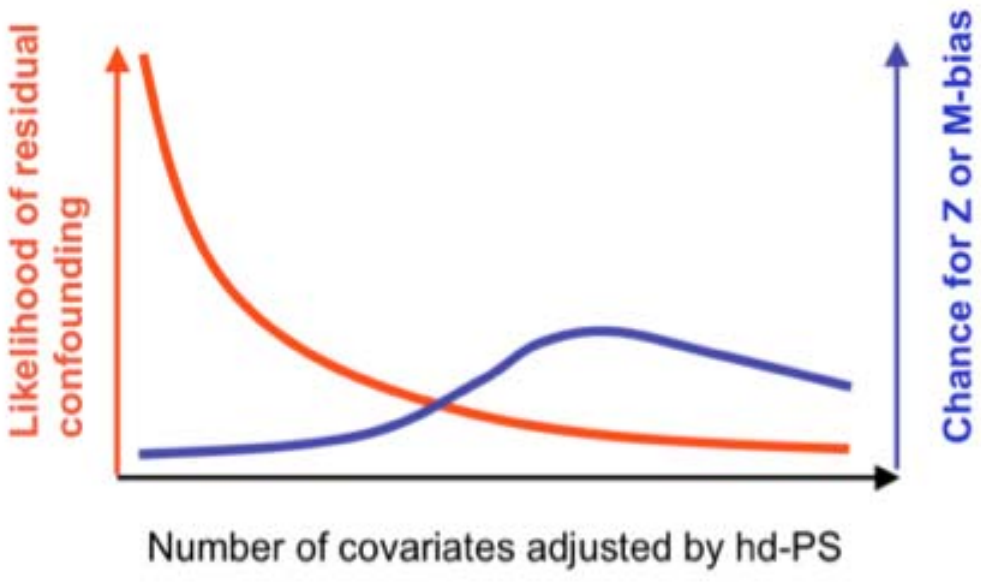


Figure 4. In most realistic scenarios, with increasing covariate adjustment net bias should be reduced even in the theoretical presence of M- or Z-bias.

4. Variable Selection Issue: Selection with Respect to Confounders

Rubin[40, 43] advocates that the process of selecting variables for a propensity score be done with no reference to the outcome; that is, variables should be chosen based on whether they balance the patients between exposure groups but not necessarily based on whether they are independent risk factors for outcome. We agree with this approach in principle, to the extent that we know which factors are confounders *a priori*. However, in real world database applications, we often do not know the entire confounder vector and, even if we were able to enumerate all theoretical confounders, we may not have knowledge of how to measure them either directly or via proxies.

A principal innovation of hd-PS is to automatically identify a large number of covariates and prioritize them according to their potential to be confounders. Contrary to Rubin's principle, this prioritization requires reference to the study outcome, but it is important to note that hd-PS algorithm checks covariate-outcome associations unconditionally and one-by-one such that the magnitude of the exposure-outcome association does not inform the variable selection process. This differs substantially from a forward or backward selection algorithm; those approaches start with a limited number of investigator-specified covariates and then select or unselect variables based on their conditional effect on the outcome. Forward and backward selection will result in falsely narrow confidence intervals, whereas we observed no such bias when we bootstrapped standard errors in hd-PS.[6] We view hd-PS as a pragmatic implementation of Rubin's principle, with reference to the outcome made in order to improve the variable selection process and thus the ultimate study validity.

5. Use of hd-PS with Few Exposures or Outcomes

In a distributed medical product active safety surveillance system, small study size issues can arise in multiple ways: it is possible that some contributing sites will be small or that sites, though large, will have low exposure frequency. The latter case may be common as new drugs are being adopted since physician unfamiliarity or restrictive formularies can limit the number of early users.

With few exposures, propensity scores are difficult to estimate. Alternatives exist—high-dimensional disease risk scores (hd-DRSs) can be used instead, with current or historical data used to estimate the score—but in an automated system, a single approach is optimal. In cases of few exposures and even fewer events, several pragmatic approaches can be contemplated: one can wait until sufficient exposures accumulate; one can pool sites to estimate propensity scores (if data sharing is allowed); one can estimate a propensity score with a minimal number of variables and accordingly lower confidence in the resulting point estimate; or one can estimate a disease risk score from historical data and use that until enough exposed patients are observed. Waiting for exposure levels to accumulate to a reasonable point may be the best choice: (1) if the overall population exposure is rare and thus the public health issue is limited; or (2) if a single data partner contributes little data so that initially excluding it may not have much effect on the pooled estimate, as little information is lost. Individual data partners with few exposed patients give rise to the concern that reimbursement rules and tight disease management programs may direct patients away from the study drug and those left using it may be highly selective, increasing concern about confounding.

With respect to few outcomes, we have shown that hd-PS works well with approximately 150 or more outcomes.[11] With between 25 and 150 outcomes, we recommend enabling the new "zero cell correction" or "confounder-exposure assessment only" options provided in the latest versions of the hd-

PS algorithm.[11] The zero cell correction adds a small number (0.1) to each cell of the confounder-outcome 2x2 table. This enables estimation of the confounder-outcome association and the variable's possible bias when these figures would otherwise have been undefined. We have shown that this option allows hd-PS to function closer to optimally when there are few outcomes, but that it can impede hd-PS's variable selection when there are more than 150 outcomes by unnecessarily forcing confounder-outcome associations toward the null. The confounder-exposure assessment-only mode judges potential for bias only via the variable's differential prevalence in the exposed and unexposed groups, irrespective of the outcome association, and works well when there are few outcomes. With fewer than 25 outcomes, we have observed that hd-PS generally works as well as investigator's specification of covariates but may not offer any improvement beyond that.

6. Automated Generation of Health Utilization Variables

Early users of the hd-PS algorithm saw surprising results that, upon close inspection, were due to the users' not having included health service intensity variables such as number of office visits or number of medications used[35] as investigator-defined variables in their analyses. Such service intensity variables are important proxies for health state through two mechanisms: (1) sicker patients see their providers more frequently, use more drugs, and may be hospitalized more often; and (2) more health care encounters lead to more opportunities for new diagnoses to be recorded and yield a better description of the patient's health state.

While in pharmacoepidemiology these types of variables often go alongside age and gender as key confounders, we recognize that an automated variable creation and selection technique should be able to account for service intensity without user intervention. To that end, we have updated the hd-PS algorithm to include automated calculation of level of service use for each of the types of health services considered (outpatient procedures, inpatient diagnoses, drugs, and so forth).[11] The hd-PS algorithm evaluates these calculated utilization variables in parallel with the variables representing the occurrence of individual codes and selects the utilization variables that appear to most bias the treatment/outcome under study. Early tests have shown that inclusion of health service intensity variables is equivalent to investigator-specification of these variables, to within 1% of the resulting point estimate.

7. Challenges to Using Propensity Scores in a Distributed Setting

A distributed system of data partners allows for contribution of many participating sites with varying levels of available information, but the robust covariate adjustment needed in most safety studies introduces certain logistical and analytic challenges. Ideally, a central site would receive contributing sites' full individual-level data, which the central site could then use flexibly. However, for patient and organizational privacy, it is generally not possible to share individual-level data and thus a better solution may be for each site to estimate an hd-PS and then share only de-identified information—anonymous identifier, exposure status, outcome status, hd-PS, and other non-identifying information—with a central facility.[36, 37] To be fully anonymous, these condensed share files cannot contain dates of services but can have relative dates (e.g., outcome date in days since exposure start). The cost of this approach is a set of limitations: each participating site must have the analytic capability to run preprogrammed code for hd-PS and relevant diagnostics, including the "Table 1s" needed to demonstrate balance; and any subgroup analyses must be specified *a priori*. Once these share files arrive in the central analytic hub, any type of propensity score analysis can be conducted, including matching stratified by data source, trimming, and regression modeling. In studies to date, including a

large-scale, multi-site investigation of the comparative safety of biologics,[24] the benefit of maximal confounding adjustment and thus maximal study validity has outweighed the limitations imposed.

However, in a distributed setting with heterogeneous data elements, hd-PS also offers two substantial advantages beyond masking patients' individual covariates. First, the algorithm is largely agnostic to data structure and coding schemes: hd-PS has worked without modification on data from Medicare, commercial US insurers, British Columbia's provincial insurance programs, and the UK's THIN and GPRD research databases. As such, it can deliver substantial covariate adjustment with very little database-specific programming or even need for transformation to a common data model. Second, hd-PS is designed to take maximal advantage of data available from each site; if one site has detailed, medical record-based information, while another has just basic claims, hd-PS will adjust maximally in each site rather than revert to a "lowest common denominator" of available information. If sites have substantially differing point estimates, we have proposed methods to determine whether the variability comes from heterogeneous patient populations or from insufficient confounding adjustment in sites with less information stored in their databases.[37]

8. Requirements for Computing Time

The original version of hd-PS was a SAS macro that required approximately 20 to 30 minutes to run on typical analyst-level desktop computer systems for cohorts of about 30,000 patients. Since the original publication of the algorithm, we have issued two major revisions, each of which substantially improves the computing time required. The first revision re-implemented hd-PS as a Java program called from within a SAS macro and reduced computing time to approximately 3 to 8 minutes, depending on available memory. This Java program takes advantage of multi-core processors by examining many covariates in parallel; the original SAS macro could only examine one covariate at a time. A second revision is targeted at high-end database machines such as IBM's Netezza database appliance; it re-implements hd-PS as a series of parallelized SQL queries with a SAS macro shell and reduces computing time to approximately 20 to 30 seconds. These improvements in computing time allow for fast, reliable execution of hd-PS and the accompanying revised algorithm architectures also allow future flexibility to implement new options and analyses, such as high-dimensional disease risk scores or scores in which the covariate-outcome association is estimated using historical data.

9. Diagnostics and Presentation of Results

Automated variable selection algorithms can appear to be somewhat of a "black box"; to counter this opacity, we have created an extensive web site (http://www.hdpharmacoepi.org) for making hd-PS's activities transparent. If the user requests, the algorithm can automatically upload and archive aggregated diagnostic information and anonymous summary data similar to a detailed "Table 1". Once it is uploaded, a public link can be generated for study investigators and external reviewers to interactively browse the variables selected, review diagnostic reports, and compare the variables selected across studies with similar exposures and outcomes. One important diagnostic available on the site is an automatically-generated Z-bias screening report that notes variables with strong exposure association but weak outcome association. Note that, although the site can display extensive information about each analysis, no individual patient data is ever in any way transmitted, visible, or inferable.

E. CONCLUSION

Active safety monitoring systems will require certain decisions to be made by investigators, while other key study elements may be automated. We feel that covariate creation and selection can be accomplished effectively with an automated process such as hd-PS or hd-DRS. We recognize the theoretical trade-offs—including variables that may be chosen because of observed associations rather than from subject matter expertise or variables that may be included unnecessarily or even incorrectly—but in studies to date, hd-PS has made choices that provide equal or better confounding adjustment compared to investigator-driven covariate selection and we do not believe that "over-adjustment" resulting in M-bias and Z-bias are threats to validity in realistic safety surveillance settings. With further study since its original publication and the resulting improvements, we feel that hd-PS is well-suited to provide automated covariate adjustment in a medical product active safety surveillance system. Settings of very few exposed patients combined with rare outcomes will remain challenging.

III. EFFECTS OF ADJUSTING FOR INSTRUMENTAL VARIABLES ON BIAS & PRECISION OF EFFECT ESTIMATES: Z-BIAS

A. INTRODUCTION

In studies of exposure effect, measured and unmeasured factors that are associated with both exposure and outcome may confound the targeted causal effect. Estimating exposure effect conditional on all confounding factors yields consistent estimates,[38-41] so determining which factors should be measured and conditioned upon is an important step for ensuring the validity of effect estimates. Ideally, this determination is made on the basis of knowledge of the relevant confounding mechanisms, but often multiple mechanisms require consideration and are not precisely known. In an attempt to mimic a randomized trial, some have argued for balancing all pre-exposure covariates between exposure groups.[42-44]

However, some literature has questioned whether instrumental variables (IVs or instruments) should be conditioned upon in effect estimation. IVs are associated with exposure but not with outcome except through their effect on exposure. IVs may be used to provide an unbiased estimate of exposure effect in the presence of unmeasured confounding via the class of IV methods. (See recent reviews[3,45-47] for precise definitions of IVs and IV methods.) Alternatively, several authors have considered the impact of conditioning on IVs in effect estimation. Rubin suggested that including an IV in the propensity score can increase the variance of an exposure effect,[24] and that result was confirmed in simulation studies.[48,49] More recently, authors have found that including IVs in the set of conditioning variables can also increase unmeasured confounding bias.[28,50,51] Bhattacharya and Vogt[50] and Patrick, et al.[52] considered empirical examples and found that including a "known" instrument in the propensity score model resulted in an estimate farther from the assumed truth than was obtained from the model that did not include the IV.

Despite the evidence of increased bias and variance of effect estimates due to conditioning on an IV, the implication of these results for epidemiologic practice remains unclear. True instruments are difficult to identify and their assumptions are not empirically verifiable.[3] For example, in a series of commentaries on a paper by Stukel, et al.,[53] authors debated whether or not the assumed IV, regional cardiac catheterization rate, was more likely to be a confounder of the association between invasive cardiac management and survival of acute myocardial infarction.[54-57] Moreover, in the presence of unmeasured confounding, IVs are statistically indistinguishable from confounders. Both types of variables will be associated with exposure and also associated with outcome conditional on exposure. Finally, the available theoretical studies of this issue provide results under a narrow range of models and do not explicate the magnitude of the increases in bias and variance. Any increase in bias due to unnecessary conditioning must be weighed against the danger of excluding real confounders from the conditioning set, an issue particularly troubling in secondary analyses of pharmacy claims data that often rely on adjusting for hundreds of confounding covariates.[7,29]

The objective of our study is to quantify the magnitude of the analytical impacts of conditioning on an IV in a range of common epidemiologic circumstances. We focus on the case where an IV may exist in the set of measured variables, but it is uncertain to investigators. We present a Monte Carlo simulation study that evaluates the effect on bias and variance under both additive and multiplicative models for outcome. We consider true instruments (variables with no direct effect on unobserved confounding

factors or outcome) and 'near-instruments'(variables that are weakly associated with the unmeasured confounding). We also explore effects, under varying assumptions, about the strength of the IV association with exposure and the magnitude of the unmeasured confounding.

B. MATERIALS AND METHODS

1. Review of the Theory

We refer to X as the exposure of interest and Y as an outcome that may be caused by X. We assume that there exists an unobserved factor, U, that confounds the association between X and Y and a measured covariate, Z. If Z satisfies the criteria of an IV for the exposure-outcome pair (X, Y), there should be no association between Z and Y except through X, as shown in **Figure 5**. We may think of this graph as representing residual associations after controlling for a vector of measured confounders. In addition, U may represent a constellation of many unobserved confounders and Z may represent the combined effect of multiple instruments. The true exposure effect and target of estimation is β2. The parameter α2 controls the strength of the IV association with exposure. The magnitude of confounding is dependent on both α1 and β1.

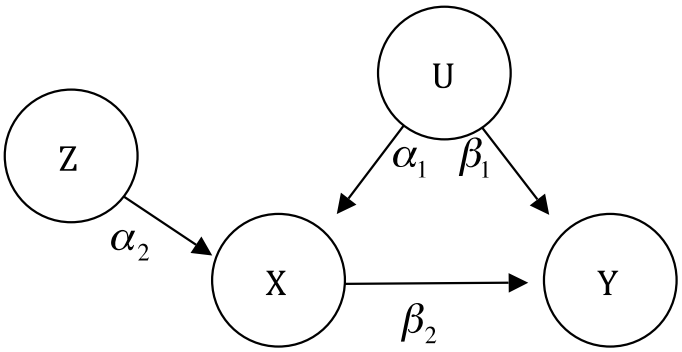


Figure 5. Causal diagram showing an unmeasured confounder, U, and instrumental variable, Z, of the exposure-outcome pair (X, Y).

We want to compare the bias of the crude, unadjusted estimator of exposure effect (given by the regression coefficient of Y on X) with the bias of the estimator for exposure effect that conditions on Z (given by the regression coefficient on X in the regression of Y on X and Z). We follow the example of Pearl[28] and assume a linear structural equation framework among zero-mean, unit-variance variables. Under these assumptions, the crude association between X and Y is given by

$$E(Y|X = x + 1) - E(Y|X = x) = \beta_2 + \alpha_1\beta_1$$

This quantity is biased for estimation of β_2 owing to confounding from U, and the bias is equal to $\alpha_1 \beta_1$. The association between X and Y conditional on Z is given by

$$E(Y|X = x + 1, Z = z) - E(Y|X = x, Z = z) = \beta_2 + \alpha_1\beta_1/(1- \alpha_2)^2$$

The bias of this estimator is $\alpha_1\beta_1/(1-\alpha_2^2)$, which is greater than $\alpha_1\ \beta_1$ whenever β_1, α_1, and α_2 are all non-zero. If α_1 or β_1 is zero, then both estimators are unbiased. If $\alpha_2 = 0$, then these biases are equal.

Therefore, in this scenario, conditioning on an IV increases the bias of the exposure effect estimator compared with the unadjusted estimator. This phenomenon can be explained intuitively if we think of partitioning the variation in the exposure variable, X, into three components: the variation explained by Z, the variation explained by U, and the unexplained variation. The proportion of the variation explained by U, along with the association between U and Y, determines the magnitude of the unobserved confounding. When we condition on Z, we effectively remove one source of variation, thereby making the variation explained by U a larger proportion of the remaining total variation in X. Thus, the residual confounding bias from U is amplified as a result of conditioning on Z. This intuition holds whenever there is unobserved confounding and an IV, regardless of the specific assumptions made above. In the Appendix, we provide an example dataset that exhibits bias amplification.

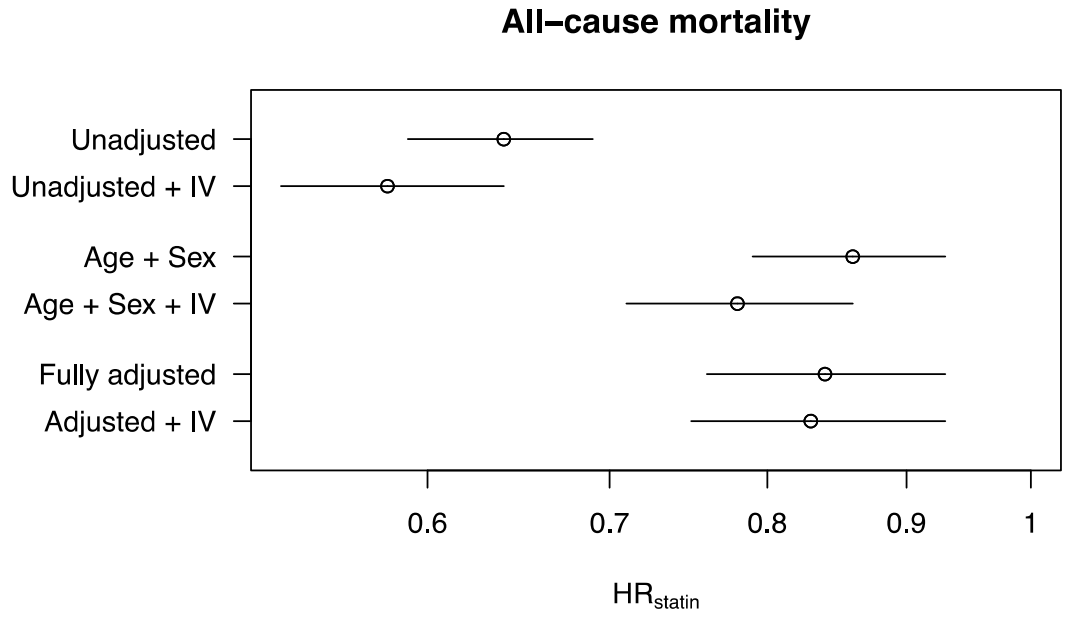
2. Empirical Example

We present the example of Patrick, et al.[52] to provide context for the simulations to follow. In this study, mortality and hip fracture were compared among elderly initiators of statin therapy and glaucoma drugs. (Source population and cohort descriptions are available in the Appendix.) Demographic characteristics, pre-treatment diagnoses, and pre-treatment health system service use were extracted to define 202 potential confounders. The investigators compared methods of selecting confounders for inclusion in the propensity score model for exposure to statins versus glaucoma drugs. The inclusion of one covariate, prior glaucoma diagnosis, resulted in effect estimates that consistently moved away from the expected effect based on the evidence from randomized controlled trials (see **Figure 6**).

Glaucoma diagnosis is strongly negatively associated with exposure to statins versus glaucoma drugs (odds ratio of 0.07) but it does not independently predict mortality or hip fracture. Therefore, glaucoma diagnosis appears to be acting as an IV in this example, since the association with exposure is much stronger than the association with outcome. Thus, the observed changes in effect estimates may be a manifestation of bias amplification. Although this study is one of the most extreme examples of bias amplification documented in the literature (an increase of 21% in the fully-adjusted analysis of hip fracture), all the estimates are so heavily biased that including the IV in the propensity score model did not alter study conclusions. In addition, the strength of the IV-exposure relation in this example made the IV easy for investigators to identify and remove.

3. Monte Carlo Simulation Studies

Pearl[28] provides formulas for the increase in bias associated with conditioning on an IV or near-IV under the assumption of zero-mean and unit-variance for all variables. To overcome these constraining assumptions, we performed two Monte Carlo simulation studies to investigate the impact on bias of conditioning on a variable that may be an IV and further considered its impact on variance. In the first experiment, we simulated data under an additive model and assumed that the goal of estimation is the risk difference in outcome between levels of exposure. In the second experiment, we simulated data under a multiplicative model and considered the goal of estimation to be the risk ratio of outcome by levels of exposure. For simplicity and to reflect a common study framework, all variables are binary.



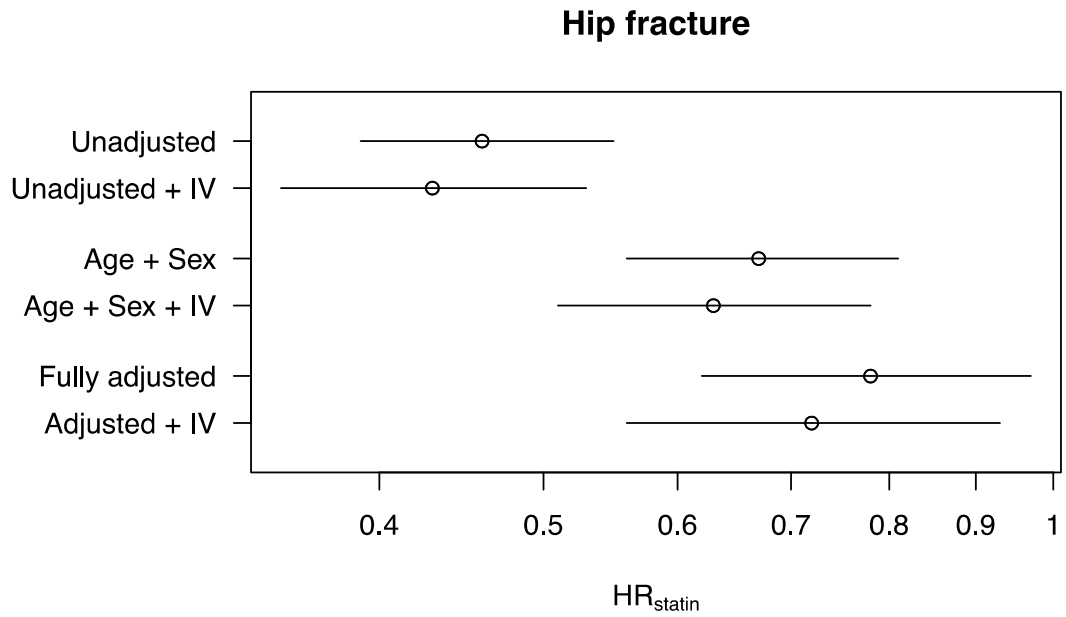


Figure 6. Estimates of the hazard ratio comparing mortality and hip fracture in statin initiators versus glaucoma drug initiators. The adjustment covariates used for each estimate are described in the left margin. The x-axis is on the log-scale with tick-marks unlogged. The approximate expected effect for mortality is a hazard ratio of 0.85. The approximate expected effect for hip fracture is a hazard ratio of 1.0.

Both simulation studies assumed the same basic causal structure, shown in **Figure 7**. The true exposure effect and target of estimation is β_2. Note that Z is not a perfect instrument in **Figure 7** as it was in

Figure 5 because it is associated with the unmeasured confounder U through γ_1. However, by varying the value of γ_1, we explored the impacts of conditioning on Z when it is a true instrument ($\gamma_1 = 0$) and when it is a near-instrument or confounder. As shown by Pearl, bias amplification may result even when the conditioning variable is not a perfect instrument.[28] In addition, we consider relatively large values of γ_1 to compare the risks of adjusting for an IV with the benefits of adjusting for a real confounder.

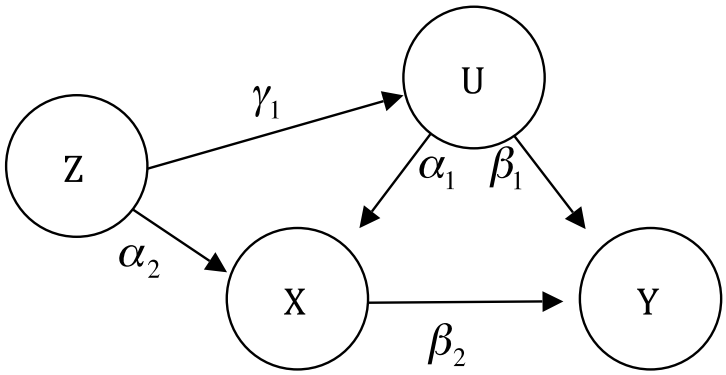


Figure 7. Causal diagram for the simulation studies.

4. Simulation Under Additive Risk

In each dataset, we simulated a binary instrument, Z, with Pr(Z = 1) = 0.5 and binary variables U, X, and Y such that

$$\Pr(U = 1|Z) = \gamma_0 + \gamma_1 Z$$

$$\Pr(X = 1|U, Z) = \alpha_0 + \alpha_1 U + \alpha_2 Z$$

$$\Pr(Y = 1|U, X) = \beta_0 + \beta_1 U + \beta_2 X$$

Variables were simulated in the above order so that the risk of outcome depends directly on U and X and indirectly on Z. The parameters γ_0, α_0, and β_0 define the baseline prevalence of each variable and each coefficient parameter may be interpreted as a conditional risk difference (RD). The values considered for each parameter are listed in **Table 2**. These values were chosen to provide the widest possible range of scenarios within the (0, 1) probability bounds for each variable. We considered two values for the baseline risk of outcome, $\beta_0 = \{0.01, 0.2\}$, corresponding to rare and prevalent outcomes, respectively. Based on the value of β_0, we constructed a range of possible values for β_1. Within this restriction, we considered all possible combinations of parameter values, resulting in 1,280 unique simulation scenarios. We included only two values for the exposure effect, β_2, because bias is invariant

to the value of this parameter. We included only positive parameter values to make the illustration of concepts as clear as possible and to avoid repeating scenarios that are symmetric and yield identical results.

For each simulation scenario, we simulated 2,500 datasets of size n =10,000. In each dataset, we calculated:

- the **crude** risk difference between X and Y : RD_{crude}
- the Mantel-Haenszel risk difference[58] between X and Y **conditional on Z**: RD_{cond}

Both RD_{crude} and RD_{cond} are estimators of the exposure effect, and we compared the performance of these two estimators.

Table 2. Parameter values for the additive simulations. The value of β0 determines the set of potential values for β1 and β2. Within that restriction, all possible combinations of parameter values were considered. The corresponding risk ratios are calculated based on the baseline prevalence of each variable and will vary depending on the values of other variables.

Variable	Baseline risks	Risk differences	Corresponding risk ratios
U	γ_0 = 0.3	γ_1 : 0, 0.006, 0.06, 0.24, 0.6	1.0, 1.02, 1.2, 1.8, 3.0
X	α_0 = 0.3	α_1: 0, 0.06, 0.18, 0.33	1.0, 1.2, 1.6, 2.1
		α_2: 0, 0.06, 0.18, 0.33	1.0, 1.2, 1.6, 2.1
Y	β_0 = 0.2	β_1: 0, 0.08, 0.36, 0.5	1.0, 1.4, 2.8, 3.5
		β_2: 0, 0.2	1.0, 2.0
	β_0 = 0.01	β_1: 0, 0.004, 0.018, 0.5	1.0, 1.4, 2.8, 51
		β_2: 0, 0.2	1.0, 21.0

5. Simulation Under Multiplicative Risk

Using the same binary instrument as in the additive study, we simulated binary variables U, X, and Y such that

$$Pr(U = 1|Z) = \gamma_0\gamma_1^{Z}$$

$$Pr(X = 1|U, Z) = \alpha_0\alpha_1^{U}\alpha_2^{Z}$$

$$Pr(Y = 1|U, X) = \beta_0\beta_1^{U}\beta_2^{X}$$

Simulating variables in the above order creates data with the causal structure depicted in **Figure 7** with associations parameterized as conditional risk ratios (RRs). The values considered for each parameter are listed in **Table 3**. We again considered all possible combinations of parameter values, resulting in 1,440 unique simulation scenarios. We used multiple values of the true exposure effect, β_2, since bias is no longer invariant to its value.

In each scenario, we simulated 2,500 datasets of size n =10,000 and calculated:

- the **crude** risk ratio between X and Y: RR_{crude}
- the Mantel-Haenszel risk ratio[62] between X and Y **conditional on Z**: RR_{cond}

As in the additive simulations, we compared the two estimators of exposure effect, RR_{crude} and RR_{cond}.

Table 3. Parameter values for the multiplicative simulations. The value of β0 determines the set of potential values for β1 and β2. Within that restriction, all possible combinations of parameter values were considered. The corresponding risk differences were calculated based on the baseline risk of each variable and will vary depending on the values of other variables.

Variable	Baseline risks	Risk ratios	Corresponding risk differences
U	γ_0 = 0.3	γ_1 : 1, 1.02, 1.2, 1.8, 3	0, 0.006, 0.06, 0.24, 0.6
X	α_0 = 0.3	α_1: 1, 1.1, 1.3, 1.8	0, 0.03, 0.09, 0.24
		α_2: 1, 1.1, 1.3, 1.8	0, 0.03, 0.09, 0.24
Y	β_0 = 0.2	β_1: 1, 1.2, 2.2	0, 0.04, 0.24
		β_2: 1, 1.2, 2.2	0, 0.04, 0.24
	β_0 = 0.01	β_1: 1, 2.2, 8.0	0, 0.012, 0.07
		β_2: 1, 2.2, 8.0	0, 0.012, 0.07

6. Evaluation of Estimator Performance

These simulation studies were designed to compare the performance of estimators for β_2 with and without conditioning on Z. For an estimator of exposure effect $\hat{\beta}_2$, we estimated the bias with the equation

$$Bias = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_2(s) - \beta_2$$

where $\hat{\beta}_2(s)$ is the value of $\hat{\beta}_2$ in the s^{th} dataset and S = 2500 is the number of simulated datasets. We estimated the standard error of $\hat{\beta}_2$ using the square-root of the sample variance of $\hat{\beta}_2(s)$ across simulated datasets. We calculated the bias and variance of the exposure effect estimators separately in each simulation scenario.

C. RESULTS

1. Additive Simulation

Figure 8 shows the performance of RD_{crude} on the x-axis versus that of RD_{cond} on the y-axis. The left panel displays the biases of both estimators; the right panel, the standard errors. Results are shown for all simulation scenarios with $\beta_0 = 0.2$, $\beta_2 = 0$, and three values of γ_1. All values of β_1, α_1, and α_2 are shown but the values of α_1 are not differentiated. Results for other values of β_0, β_2, and γ_1 are similar to the results shown here and are available in Appendix C. In each plot, the solid diagonal marks equality. A point on the line indicates a simulation scenario where the bias or standard error is invariant to conditioning on Z; scenarios where the bias or standard error is increased or decreased by conditioning on Z are represented by points above or below the line, respectively.

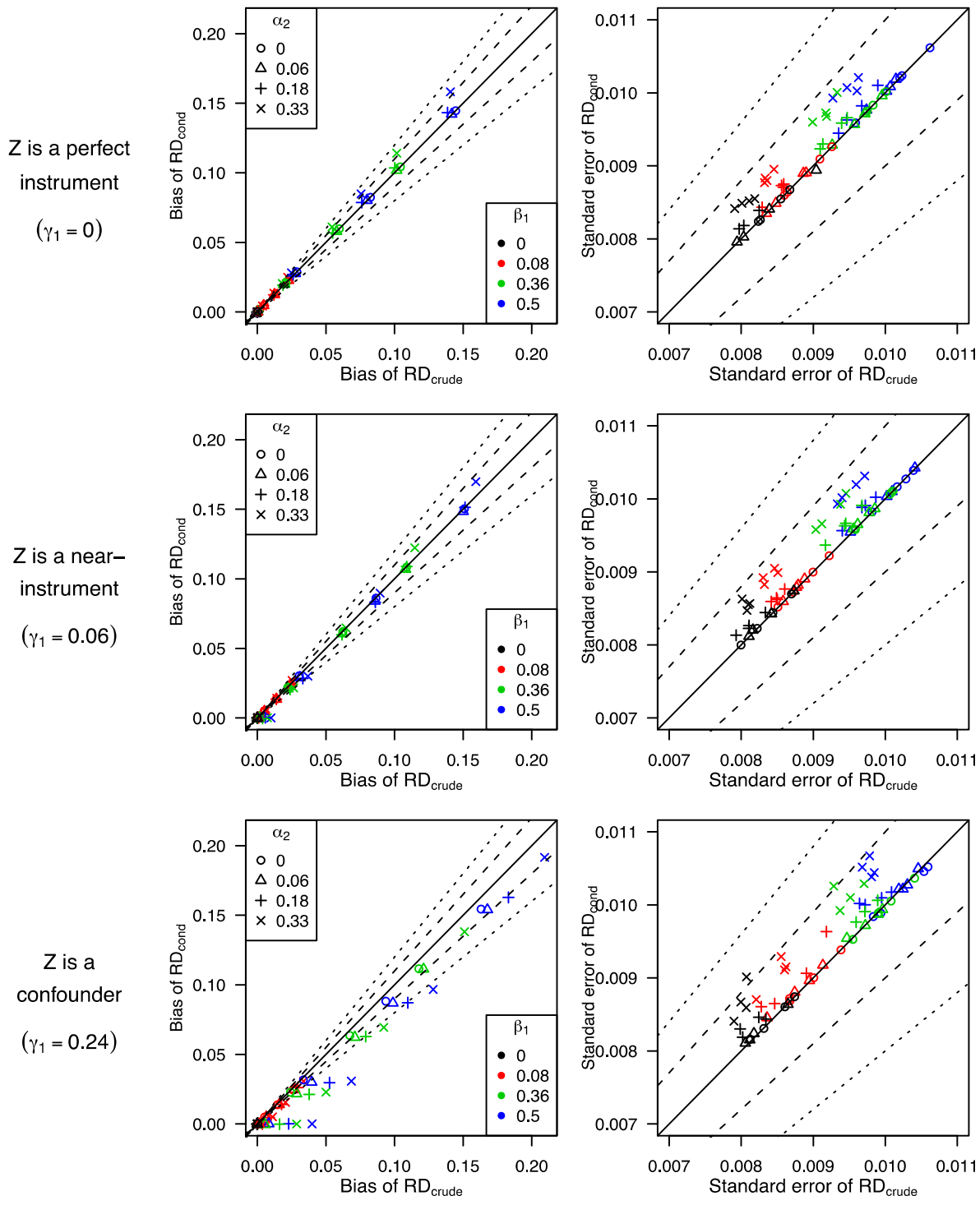


Figure 8. The bias and standard error of exposure effect estimators with and without conditioning on Z. Each point represents one simulation scenario in the additive simulations. The solid diagonal marks equality. Dashed lines mark the threshold for a 10% increase or decrease and dotted lines mark a 20% increase or decrease.

In the top row of plots in **Figure 8**, $\gamma_1 = 0$, indicating that Z is simulated to be a perfect instrument for the exposure-outcome pair (X,Y). Therefore, the bias in RD_{crude} is due to unobserved confounding from U. As predicted, conditioning on the instrument, Z, results in an estimator of exposure effect that is more biased than the crude estimator. In addition, the standard error of RD_{cond} is larger than the standard error of RD_{crude}. The magnitude of these increases depends on the value of α_2. When Z is a strong instrument ($\alpha_2 = 0.33$), the increases in bias and standard error due to conditioning on Z are largest; when Z is a weak instrument ($\alpha_2 = \{0.06, 0.18\}$), the increases are negligible; when Z has no association with exposure ($\alpha_2 = 0$), there is no increase in either bias or standard error.

In the center row of **Figure 8**, $\gamma_1 = 0.06$, indicating that Z is not a perfect instrument because Z is associated with Y through the unobserved confounder, U. However, we may consider Z to be a near-instrument (or near-confounder) because its association with U is relatively weak. **Figure 8** shows that when Z is a near-instrument, conditioning on Z may increase or decrease the bias in exposure effect estimation, depending on the values of other parameters. In particular, conditioning on Z tended to result in increased bias in simulation scenarios with the largest crude bias and decreased bias in simulation scenarios with smaller crude bias. In the former case, the unobserved confounding due to U overwhelms the relatively small amount of confounding due to Z. In the latter case, the confounding due to U is smaller and Z accounts for more of the overall confounding bias of exposure effect.

In the bottom row of **Figure 8**, $\gamma_1 = 0.24$, indicating that Z is a confounder in these scenarios. When we condition on the confounder, the bias is always equivalent or decreased but the standard error may increase or decrease. As before, the magnitude of the increase in standard error is determined by the value of α_2 with the largest increases occurring when $\alpha_2 = 0.33$.

Across all the additive simulation scenarios defined in Table 1, the largest absolute increase in bias due to conditioning on Z was an increase of 0.018 on a crude bias of 0.141. This scenario had the highest value considered for each of α_1, β_1, and α_2, and $\gamma_1 = 0$. (Equal biases were found across values of β_0 and β_2.) The largest observed increase in standard error due to conditioning on Z was an increase of 0.003 on a crude standard error of 0.009. This scenario had the highest value considered for all parameters.

Because α_2 is shown to be the most important parameter in determining the magnitude of the increases in bias and variance when conditioning on Z, we further considered a scenario with a larger value for α_2. In the case of a binary exposure, the value of α_2 is constrained by the (0, 1) bounds on probability of exposure. Therefore, in order to increase α_2, we reduced the baseline prevalence of exposure ($\alpha_0 = 0.1$) and chose the other parameter values $\gamma_0 = 0.3$, $\gamma_1 = 0$, $\alpha_2 = 0.6$, $\beta_0 = 0.2$, $\beta_1 = 0.5$, and $\beta_2 = 0$. Simulating with these values yields biases of 0.101 and 0.158 for RD_{crude} and RD_{cond}, respectively, representing a 56% increase in bias. The standard errors of RD_{crude} and RD_{cond} were 0.01 and 0.012, respectively, representing a 20% increase in standard error.

We further repeated one simulation scenario under varying study sizes to explore the bias-variance trade-off with respect to conditioning on an IV. In particular, we use the scenario reported above with the largest absolute increase in bias due to conditioning on Z from Table 1. **Figure 9** displays the standard error of RD_{crude} and RD_{cond} under a range of study sizes. The standard error increases rapidly as the study size decreases, and the increase in standard error attributable to conditioning on Z is

negligible compared with the impact of study size. Also, even at the smallest study size considered (n = 100), the standard error of both RD_{crude} and RD_{cond} is smaller than the bias in this scenario.

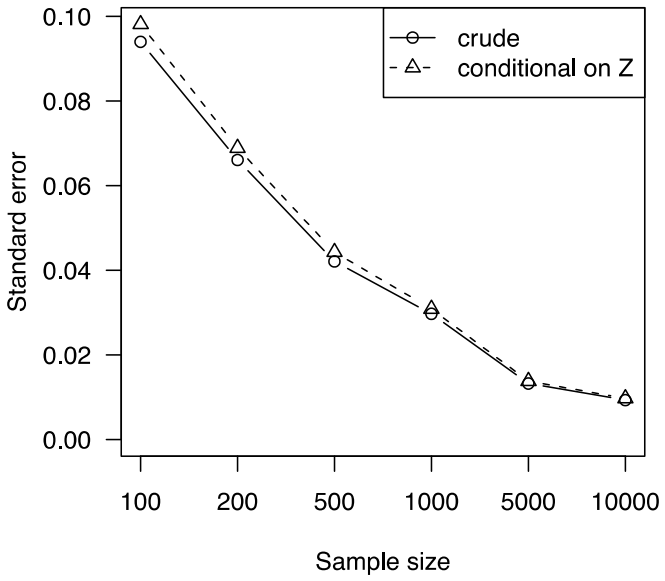


Figure 9. The standard error of exposure effect estimators with and without conditioning on Z under a range of study sizes.

2. Multiplicative Simulation

Figure 10 shows the bias (left) and standard error (right) of RR_{crude} on the x-axis versus that of RR_{cond} on the y-axis. As in Figure 4, the y = x line is provided. Results are displayed for all simulation scenarios with $\beta_0 = 0.2$, $\beta_2 = 1$, and three values of γ_1. Results for other values of β_0, β_2, and γ_1 are similar to the results shown here and are available in Appendix C. The values of the baseline risk for outcome, β_0, and the true exposure effect, β_2, determine the scale of the biases but not their relative magnitudes.

In the multiplicative simulations, associations are parameterized as risk ratios, so the three values of γ_1 shown in Figure 6 indicate that the variable Z is simulated to be a perfect instrument, a near-instrument (or near-confounder), and a confounder, respectively, for the exposure-outcome pair (X,Y). Results are similar to the results from the additive simulations. In the presence of unobserved confounding, conditioning on a perfect instrument increases the bias and standard error in exposure effect estimation and this increase tends to be larger when the instrument is strong ($\alpha 2 = 0.33$) and the crude bias or standard error is large. In the scenarios with no confounding bias from U ($\alpha 1 = 1$ or $\beta 1 = 1$), conditioning on Z does not create bias. Conditioning on a near-instrument tends to result in increased bias when the crude bias is large and in decreased bias when the crude bias is relatively small. When Z is a confounder, bias generally decreases as a result of conditioning on Z but standard error may increase or decrease.

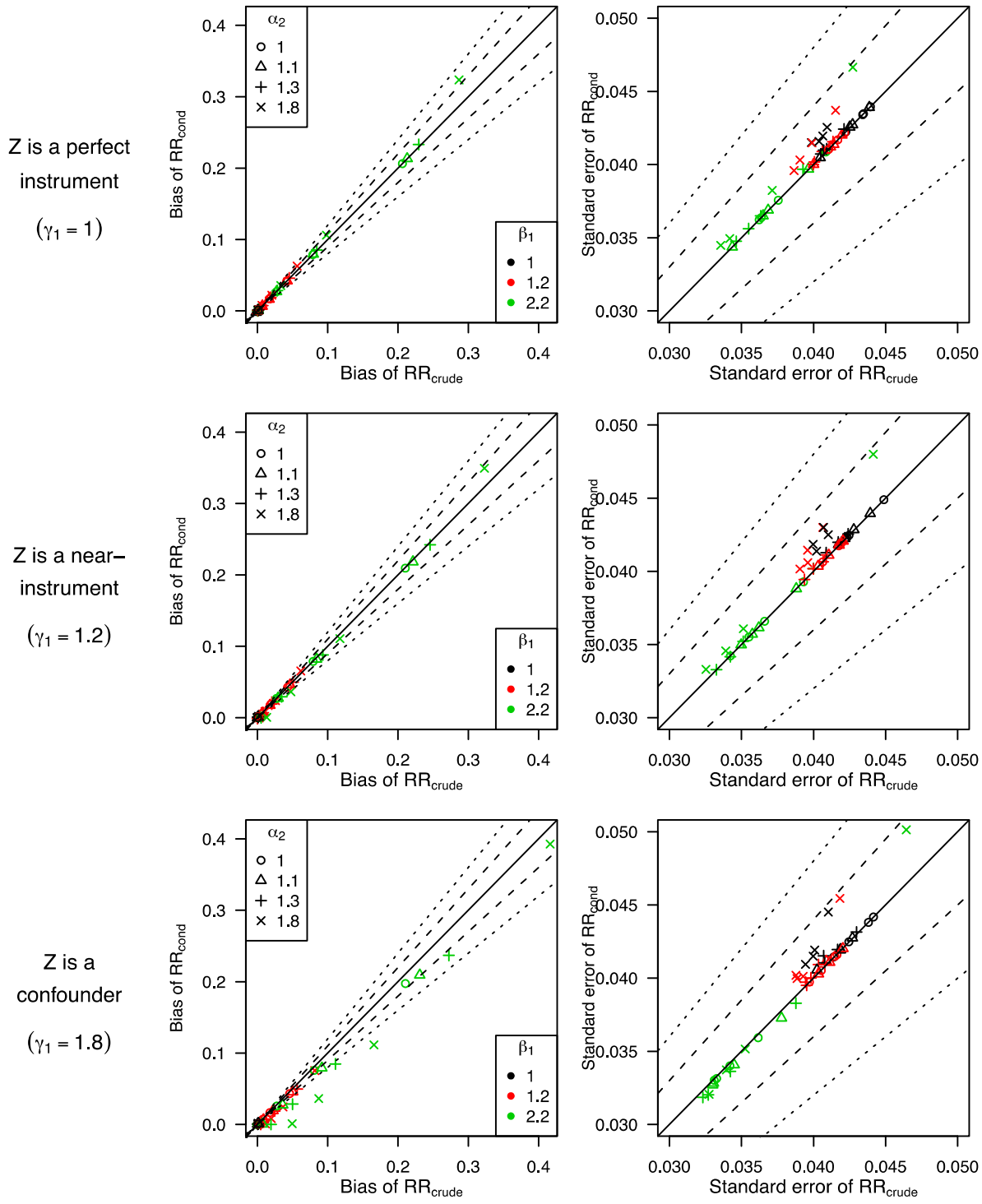


Figure 10. The bias and standard error of exposure effect estimators with and without conditioning on Z. Each point represents one simulation scenario in the multiplicative simulations. The solid diagonal marks equality. Dashed lines mark the threshold for a 10% increase or decrease and dotted lines mark a 20% increase or decrease.

D. DISCUSSION

The simulation studies showed that estimating exposure effect conditional on a true instrument can increase (and never decreases) the bias and standard error but these increases are generally small. In particular, when the unobserved confounding was small, the increase in bias and variance due to conditioning on an IV was essentially negligible. When the unmeasured confounding bias was large, the increase in estimation error due to conditioning on an IV represented only a small fraction of the overall error in most scenarios. In addition, the effects on bias of conditioning on an IV were reduced by reducing the strength of the unmeasured confounding. Based on these results, we believe that the most important task for an investigator is to minimize the unmeasured confounding, even at the risk of including IVs in the set of conditioning variables. If the IV-exposure association is very strong, as in the example study of statins and glaucoma drugs, then the IV is likely to be identified and should be removed from analysis. If the IV- exposure association is weak, then mistakenly including the IV in the set of conditioning variables will be unlikely to make a meaningful difference in the conclusions.

Increases in bias and standard error were also observed when conditioning on a measured variable that is not a true instrument when that variable is weakly associated with outcome and strongly associated with exposure. The increase in bias occurs only when the unmeasured confounding bias is larger than the confounding due to the measured variable. When conditioning on a confounder, standard error may still increase, but never bias. These results are consistent with past theoretical and simulation findings.[28,48,50,51,59] When comparing these results with studies that use continuous variables, we must scale effect sizes to reflect the alternate units of measurement. Results are also consistent between the two simulation experiments explored in this paper.

Although we were able to deduce consistent trends across simulation scenarios, specific findings are dependent on the specification of the data-generating process and the parameter values considered. In particular, it is clear that the magnitude of the increase in bias is limited only by the strength of the association between the IV and exposure. In the case of a binary exposure, as we used here, the strength of this association is constrained by the baseline prevalence of exposure and the effects of other factors that determine exposure. When analyzing a continuous exposure, no such constraints exist and the IV association with exposure may be larger. In addition, in cases of a known IV (for example, randomized assignment to exposure) the association between the IV and exposure may be stronger. In the simulation studies, the parameter values were chosen to represent the range of associations most likely to be encountered in epidemiologic studies with a binary exposure and a binary covariate that is not known to be an IV. Within this range, the maximum increase in bias observed in any scenario was approximately a 20% increase over the crude bias. When we further considered a scenario with a stronger association between the IV and exposure, we observed a 56% increase in bias. However, achieving this magnitude of bias increase required both extremely large unmeasured confounding as well as a very strong instrument. On the other hand, when conditioning on a confounder, a 50% or more decrease in bias was relatively easy to achieve and did not require such an extreme scenario.

Therefore, we conclude that in the case of secondary analyses of non-experimental data, for example large databases of healthcare claims, inadvertently including IVs in the set of conditioning covariates does not appear to pose a major threat to the validity of exposure effect estimates. However, these results do not imply that investigators are free from thinking about the associations among measured variables. Variables that are known to be instruments should not be conditioned upon. The belief that balancing all pre-exposure covariates, as in randomized studies, can do no harm does not hold in non-

experimental studies because unobserved factors may exist that cannot be balanced. As shown in this paper, balancing some factors can amplify the effects of factors that remain unbalanced. In addition, although instruments and confounders are empirically indistinguishable, ordering variables based on the magnitude of their association with the outcome could provide a reasonable approach to selecting covariates for conditioning, as recommended by Hill[60] and implemented in a high-dimensional propensity score algorithm.[7] Other outcome-based approaches to variable selection, such as Bayesian propensity scores,[61] may offer similar benefits.

IV. REFERENCES

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999; 10: 37-48.

2. Miettinen O, Wang J. An alterantive to the proportional mortality ratio. *American Journal of Epidemiology*. 1981; 114: 144-8.

3. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006; 17: 360-72.

4. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*. 1991; 133: 144-53.

5. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*. 2003; 158: 915-20.

6. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf*. 2010; 19: 858-68.

7. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009; 20: 512-22.

8. Rassen JA, Choudhry NK, Avorn J, Schneeweiss S. Cardiovascular outcomes and mortality in patients using clopidogrel with proton pump inhibitors after percutaneous coronary intervention or acute coronary syndrome. *Circulation*. 2009; 120: 2322-9.

9. Schneeweiss S, Patrick AR, Solomon D, et al. The comparative safety of antidepressant agents in children regarding suicide attempts and suicides. *Pediatrics*. 2010; 125: 876-88.

10. Schneeweiss S, Patrick AR, Solomon D, et al. Variation in the risk of suicide attempts and completed suicides by antidepressant agent in adults: A propensity score-adjusted analysis of 9 years of data. *Arch Gen Psych*. 2010; 67: 497-506.

11. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Observed performance of high-dimensional propensity score analyses of treatment effects in small samples. *American Journal of Epidemiology*. 2011. [In press]

12. Patorno E, Bohn RL, Wahl PM, et al. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. *JAMA*. 2010; 303: 1401-9.

13. Huybrechts KF, Rothman KJ, Silliman RA, Brookhart MA, Schneeweiss S. Risk of Death and Hospitalization for Major Medical Events after Initiation of Psychotropic Medications in Older Adults Admitted to Nursing Homes. *Can Med Assoc J*. 2011. [In press]

14. Toh S, García Rodríguez LA, Hernan MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: An application to electronic medical records. Pharmacoepidemiol Drug Saf. 2011; 8: 849-57.

15. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med*. 2000; 343: 1520-8, 2 p following 8.

16. Silverstein FE, Faich G, Goldstein JL, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA*. 2000; 284: 1247-55.

17. HD Pharamcoepi Web Site. Available at: http://www.hdpharmacoepi.org. Accessed 8/9/2011.

18. Kloss S. Propensity score & high-dimensional propensity score methods in observational studies based on administrative data of statutory health insurances. *Universität Bremen*. 2010;Master's Thesis.

19. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966; 19.

20. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. When patients are treated contrary to prediction - implications for use of propensity scores in extreme cases [abstract]. *Pharmacoepidemiol Drug Saf*. 2007; 16(suppl 2): S3.

21. Mahalanobis PC. On the generalized distance in statistics. *Proc Natl Inst Sci (India)*. 1936; 12: 49-55.

22. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: in 25 variations, the physician prescribing preference generally was strong and reduced imbalance. *J Clin Epidemiol*. 2009; 62: 1233-41.

23. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007; 26: 20-36.

24. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 1997; 127: 757-63.

25. Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B, Mukhopadhyay P. Variable selection and raking in propensity scoring. *Stat Med*. 2007; 26: 1022-33.

26. Weinberg CR. Toward a clearer definition of confounding. *American Journal of Epidemiology*. 1993; 137: 1-8.

27. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003; 14: 300-6.

28. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR: Citeseer, 2010; 425-32.

29. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care*. 2010; 48: S114-20.

30. Liu W, Brookhart MA, Setoguchi S. Impact of collider-stratification bias (M-bias) in pharmacoepidemiologic studies: a simulation study. *Pharmacoepidemiol Drug Saf*. 2010; 19: S212.

31. Angrist JD, Imbens G, Rubin DB. Identification of causal effects using instrumental variables. *JASA*. 1996; 94: 444-55.

32. Rassen JA, Brookhart MA, Mittleman MA, Glynn RJ, Schneeweiss S. Instrumental variables I: exploiting quasi-random treatment choice to construe causal relationships. *J Clin Epidemiol*. 2009; 62: 1226-32.

33. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*. 2010;19:537-554.

34. Myers, JA, Rassen, JA, Gagne, J, Huybrechts, K, Rothman, KJ, Schneeweiss, S, Joffe, MM, Glynn, RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. In press at American Journal of Epidemiology.

35. Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *American Journal of Epidemiology*. 2001; 154: 854-64.

36. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf*. 2010; 19: 848-57.

37. Rassen JA, Solomon DH, Curtis LH, Herrington L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care*. 2010; 48: S83-9.

38. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968; 24: 295-313.

39. Billewicz WZ. The efficiency of matched samples: An emperical investigation. *Biometrics*. 1965; 21: 623-43.

40. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *BIometrika*. 1983; 70: 41-55.

41. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79: 516-24.

42. D'Agostino, Jr. RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998; 17: 2265-81.

43. Rosenbaum PR. *Observational Studies*. 2nd ed. New York: Springer Verlag, 2002.

44. Rubin DB. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Stat Med*. 2009; 28: 1420-3.

45. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology*. 2006; 17: 260-7.

46. Glymour MM. Natural experiments and instrumental variable analyses in social epidemiology. *Methods in Social Epidemiology*. 2006; Sect 429-68.

47. Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Services and Outcomes Research Methodology*. 2007; 7: 159-79.

48. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American Journal of Epidemiology*. 2006; 163: 1149-56.

49. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007; 26: 734-53.

50. Bhattacharya J, Vogt WB. Do instrumental variables belong in propensity scores? *NBER Working Paper*. 2007.

51. Wooldridge J. Should instrumental variables be used as matching variables. Michigan State University. 2009. Available at: https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf. Accessed 5/30/11.

52. Patrick AR, Scheeweiss SM, Brookhart MA, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology - an empirical illustration. *Pharmacoepidemiology and Drug Safety*. 2011; 20(6): 551-9.

53. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. 2007; 297: 278-85.

54. Novikov I, Kalter-Leibovici O. Analytic approaches to observational studies with treatment selection bias. *Journal of the American Medical Association*. 2007; 297: 2077.

55. D'Agostino Jr RB, D'Agostino Sr RB. Estimating treatment effects using observational data. *Journal of the American Medical Association*. 2007; 297: 314-6.

56. Stukel TA, Fisher ES, Wennberg DE. Analytic Approaches to Observational Studies With Treatment Selection Bias-Reply. *Journal of the American Medical Association*. 2007; 297: 2078.

57. Stukel TA, Fisher ES, Wennberg DE. Using Observational Data to Estimate Treatment Effects. *Journal of the American Medical Association*. 2007; 297: 2078.

58. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008.

59. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996; 52: 249-64.

60. Hill J. Discussion of research using propensity-score matching: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in Medicine. *Stat Med*. 2008; 27: 2055-61.

61. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Stat Med*. 2009; 28: 94-112.

V. APPENDICES

A. APPENDIX A: EXAMPLE DATASET

We present an example dataset from the multiplicative simulation study, where amplification of bias and standard error were relatively large. This data was simulated under the following true parameter values: $\beta_0 = 0.01$, $\beta_1 = 8$, $\beta_2 = 8$ (the true exposure effect), $\alpha_0 = 0.3$, $\alpha_1 = 1.8$, $\alpha_2 = 1.8$, $\gamma_0 = 0.3$, and $\gamma_1 = 1$. The simulated data for one set of 10,000 patients is given in **Table 4**. The crude estimate of exposure effect from this data is $RR_{crude} = 15.52$. Thus, the bias of RR_{crude} is $15.52 - 8 = 7.52$. The estimate of exposure effect conditional on Z is $RR_{cond} = 17.07$ and the bias of RR_{cond} is $17.07 - 8 = 9.07$.

Table 4. One simulated dataset from the multiplicative simulations.

	Z = 0		Z = 1	
	Y = 1	Y = 0	Y = 1	Y = 0
X = 1	603	1258	1084	2263
X = 0	80	3059	20	1633

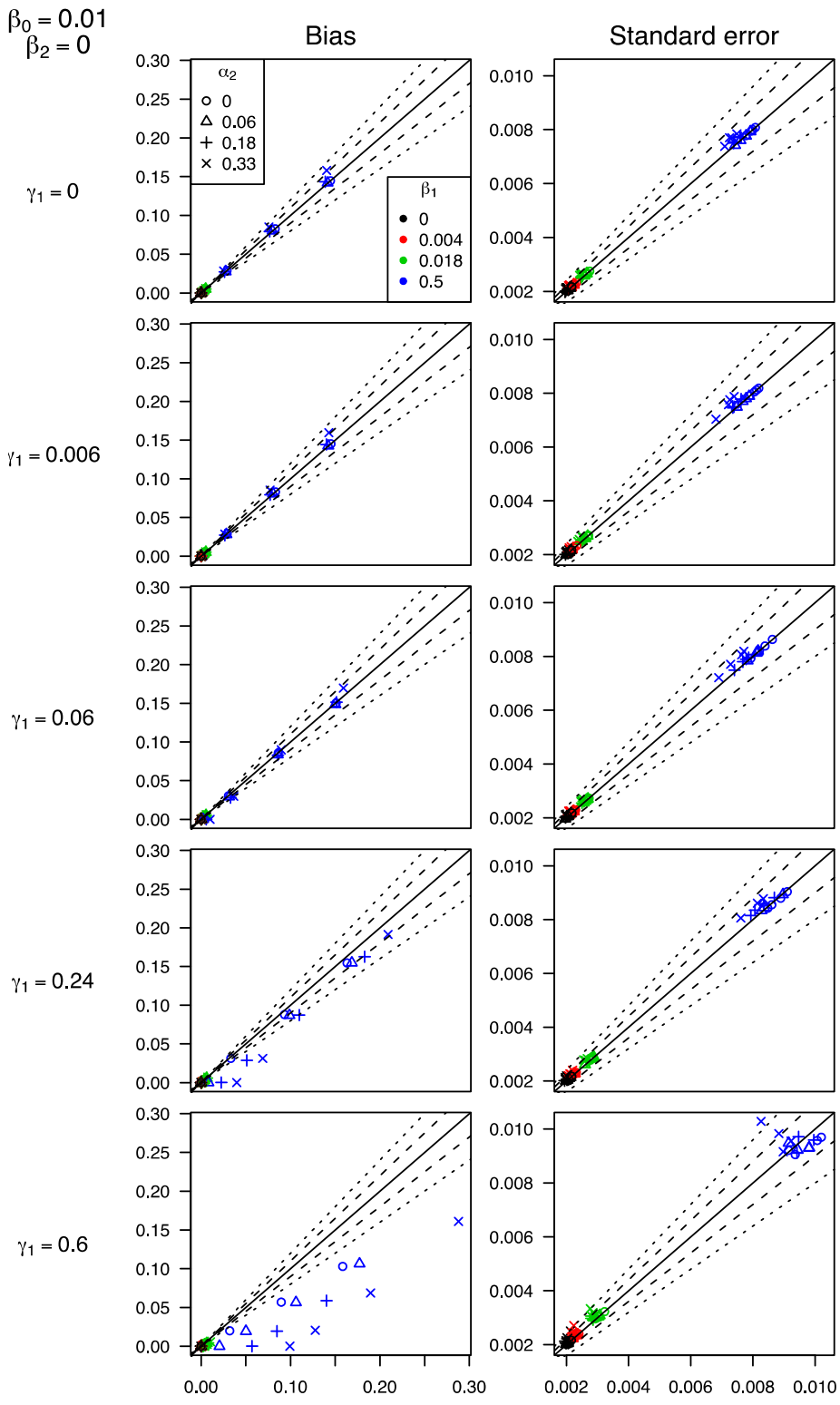
B. APPENDIX B: EMPIRICAL EXAMPLE

The data in the empirical example in Section 2 come from the investigation described by Patrick, et al.[52] This cohort study includes patients initiating statins and glaucoma drugs among Medicare beneficiaries 65 years of age and older who were enrolled in the Pharmaceutical Assistance Contract for the Elderly (PACE) program provided by the state of Pennsylvania. Enrollees in PACE were eligible for inclusion in the study population if they filled a prescription for any statin or glaucoma drug between January 1, 1996, and December 31, 2002, and demonstrated continuous healthcare system use.

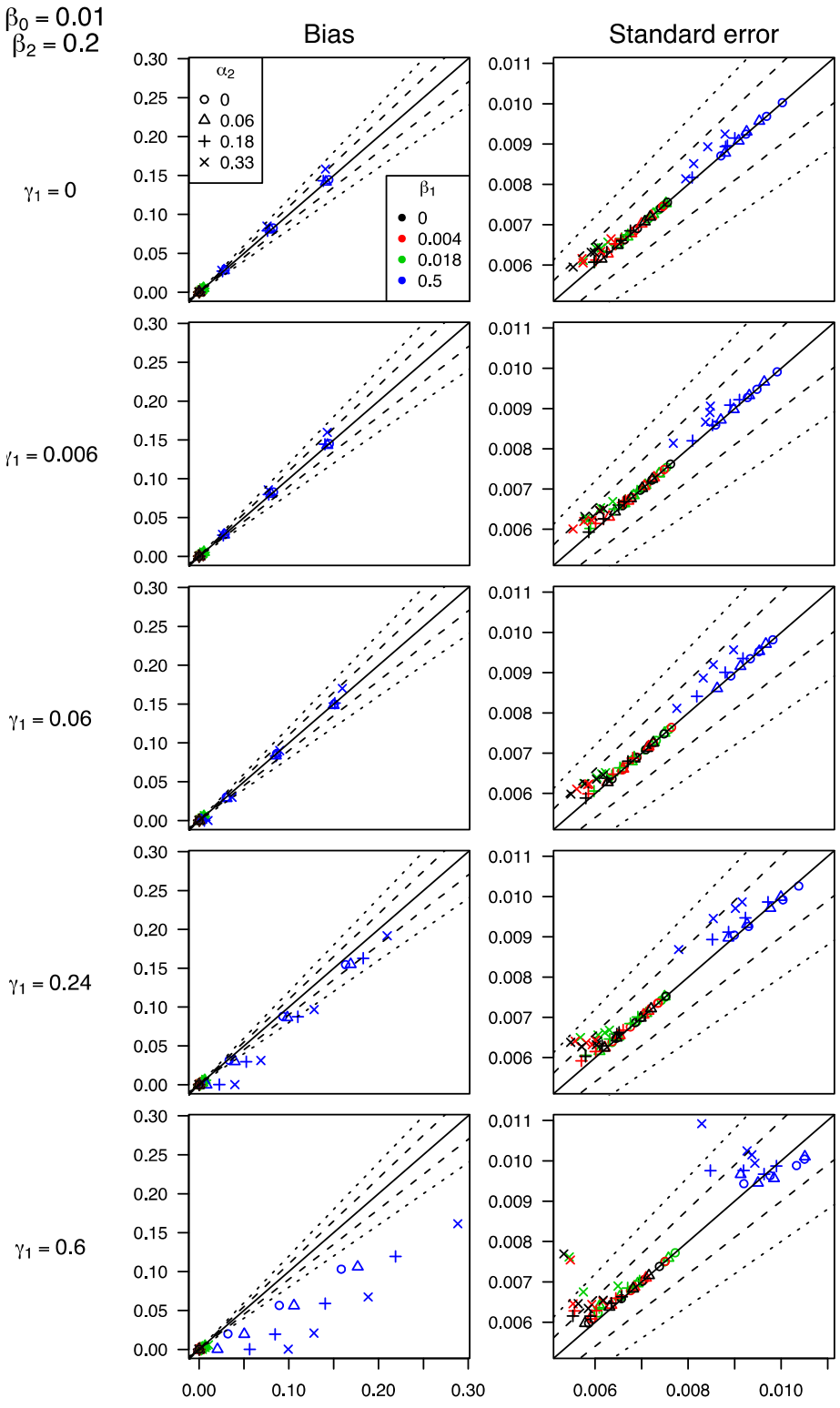
Initiation of statin use was defined as when an eligible beneficiary filled at least one prescription for a drug of interest between January 1, 1996, and December 31, 2002, but did not use one during the 18 months prior to the index date. The index date was the first date a prescription for a statin or glaucoma drug was filled. Follow-up was for one year after the initiation of therapy. Covariates were defined based on enrollment information (age, sex, race) and claims in the year before initiation of therapy.

C. APPENDIX C: SIMULATION RESULTS

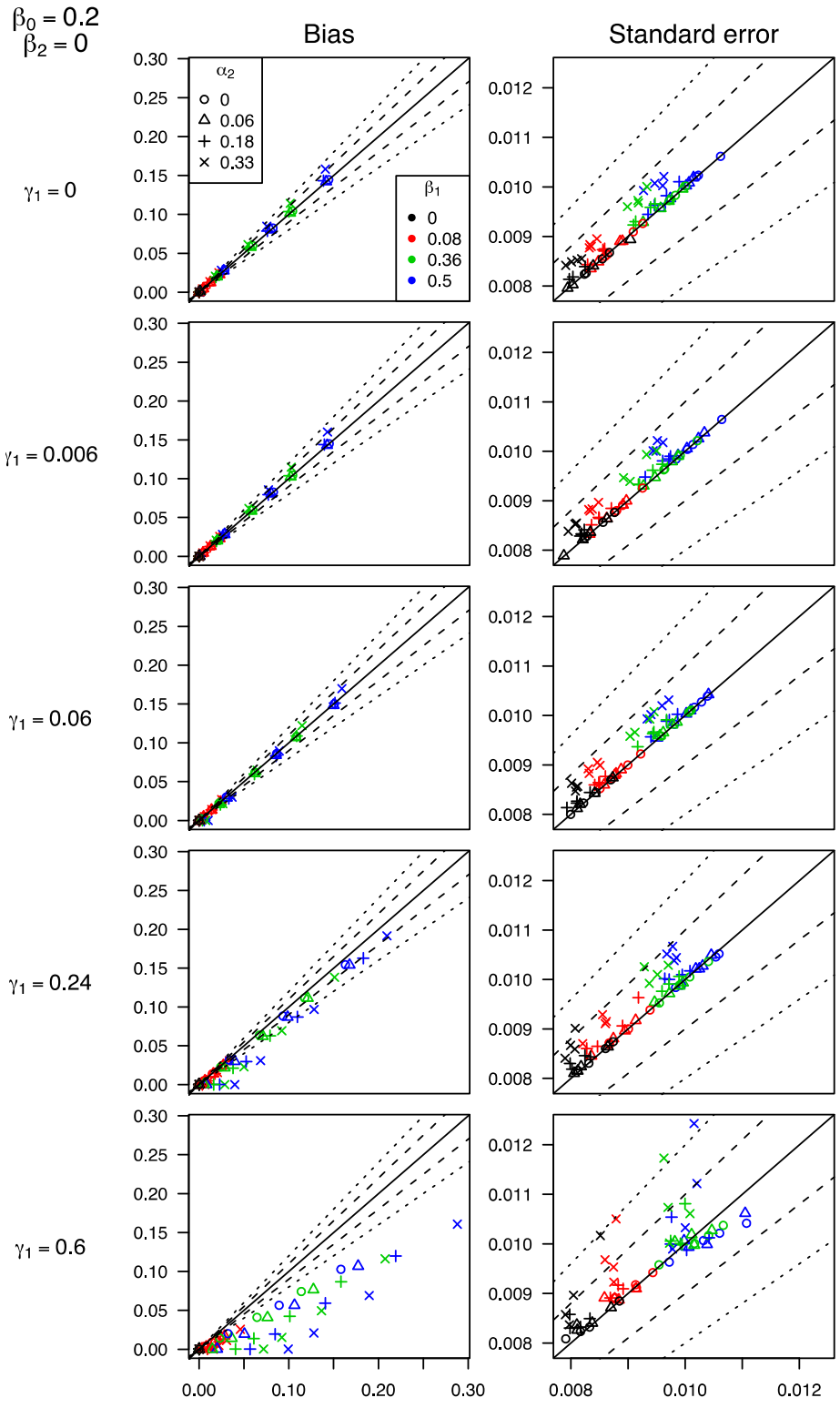
All results of both additive and multiplicative simulation studies are presented here. The figures in this section are similar to Figures 8 and 10 in the report. On the x-axis, we plot the bias (left panel) and standard error (right panel) of RD_{crude}. On the y-axis, we plot the bias and standard error of RD_{cond}. Each page contains all scenarios for a unique combination of the values of β0 and β2 and these values are marked in the top left corner of each page. Each row of plots further distinguishes the values of γ1, marked to the left of each row. Within each plot, results for all values of α1, α2, and β1 are presented, but the values of α1 are not differentiated. The solid diagonal marks equality. Dashed lines represent a 10% increase or decrease, and dotted lines represent a 20% increase or decrease. Appendix Figures 1-4 are from the additive simulations, and Appendix Figures 5-10 are from the multiplicative simulations.



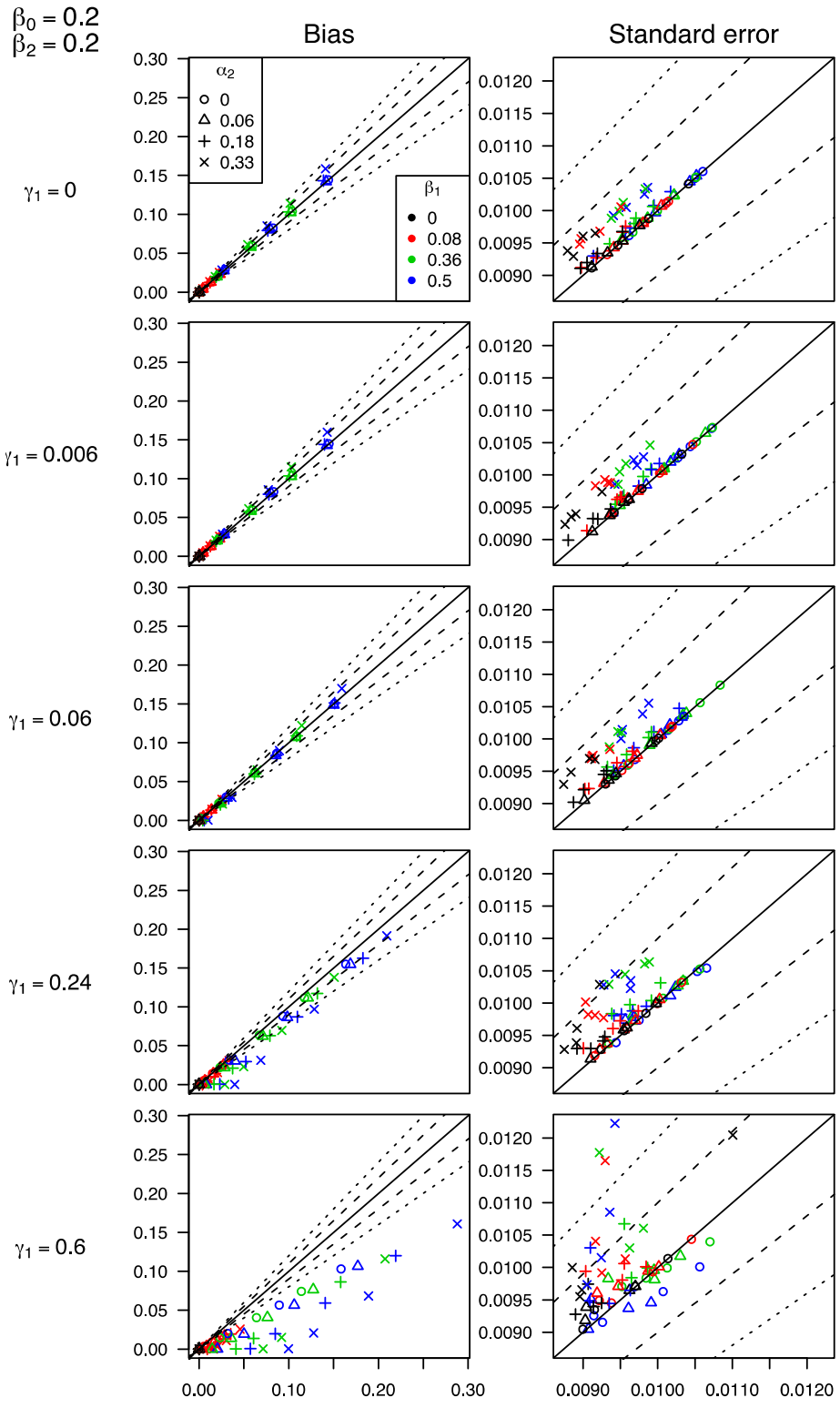
Appendix Figure 1: Each point represents one simulation scenario in the additive simulations with β_0=0.01 and β_2=0.



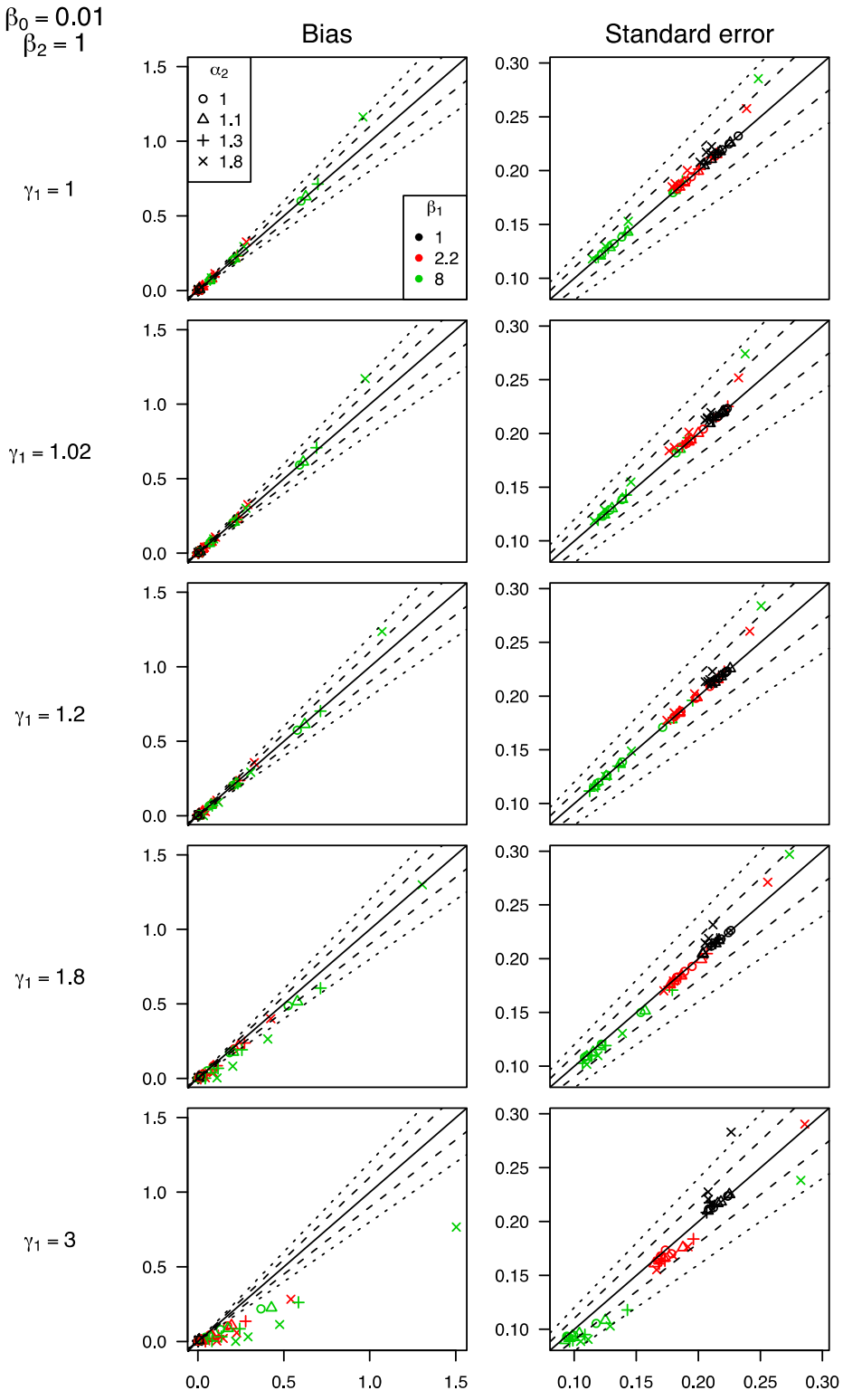
Appendix Figure 2: Each point represents one simulation scenario in the additive simulations with β_0=0.01 and β_2=0.2.



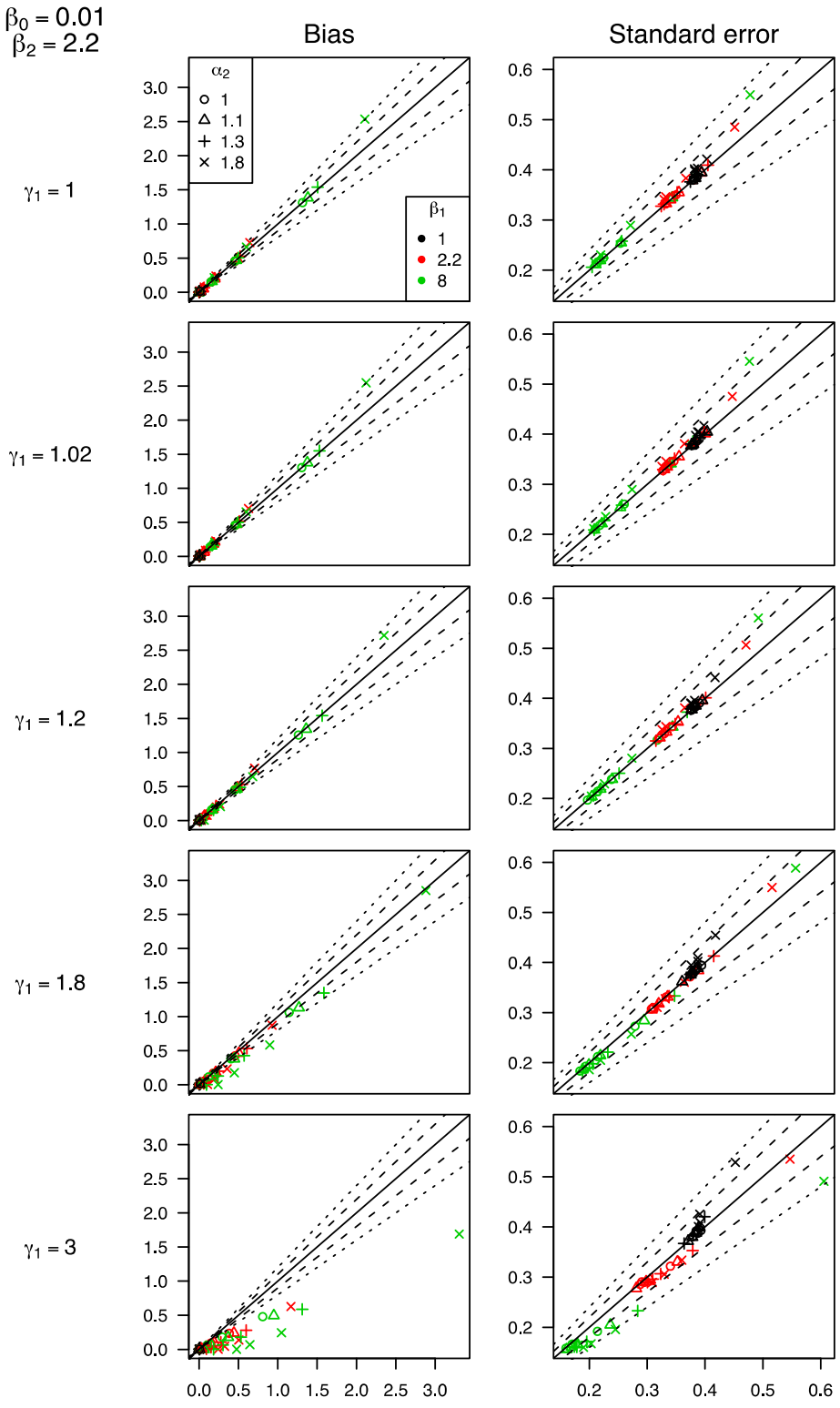
Appendix Figure 3: Each point represents one simulation scenario in the additive simulations with β_0=0.2 and β_2=0.



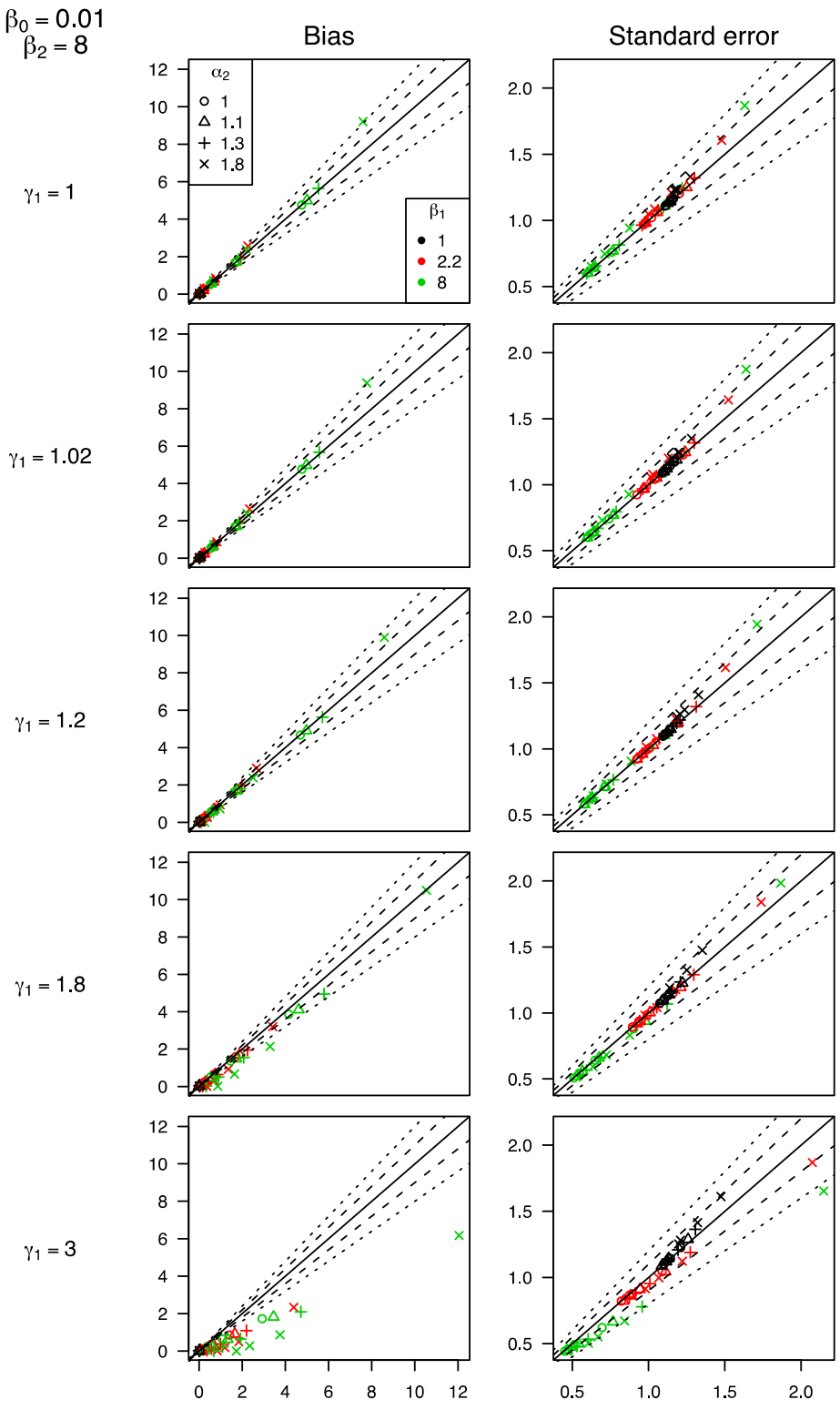
Appendix Figure 4: Each point represents one simulation scenario in the additive simulations with β_0=0.2 and β_2=0.2.



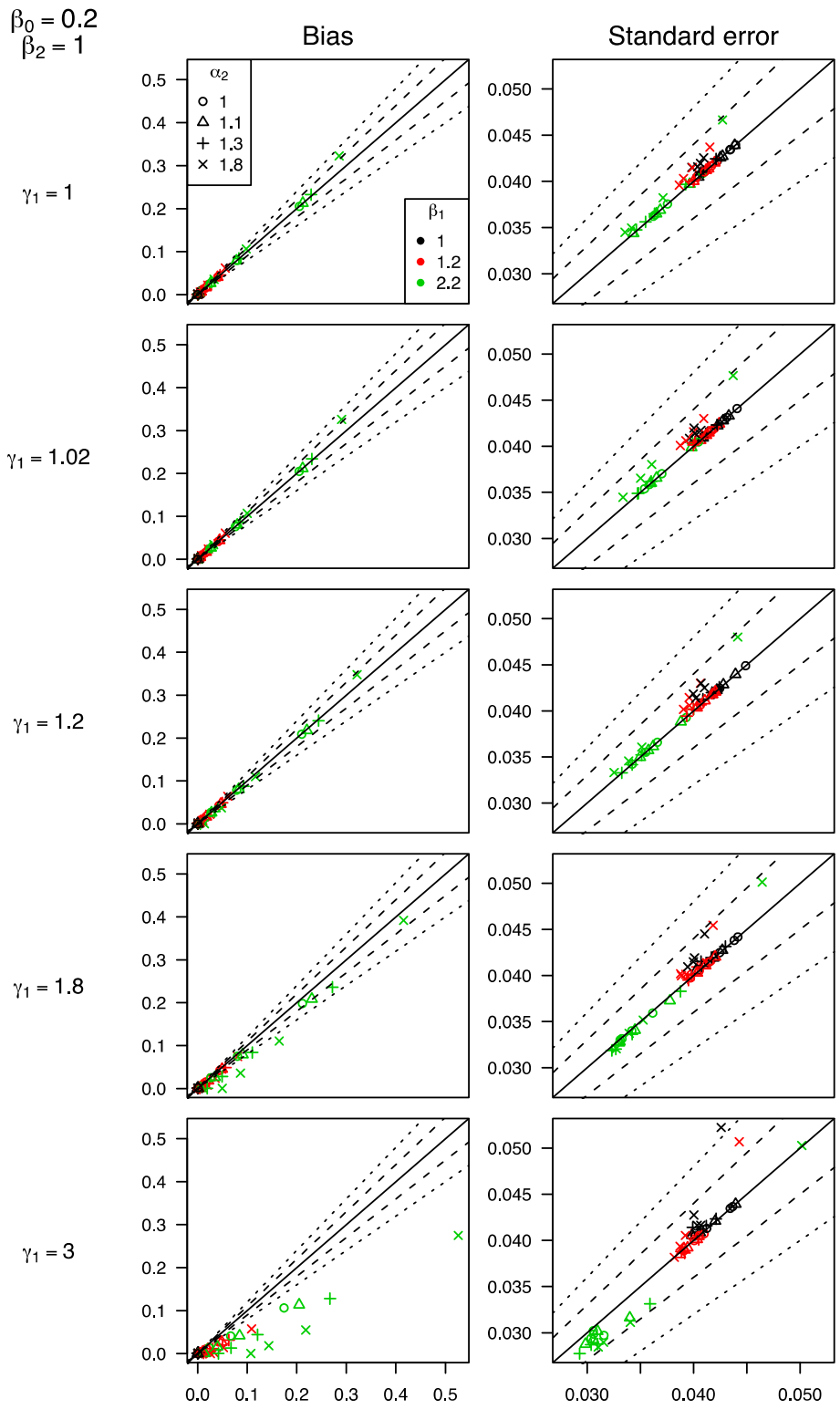
Appendix Figure 5: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.01 and β_2=1.



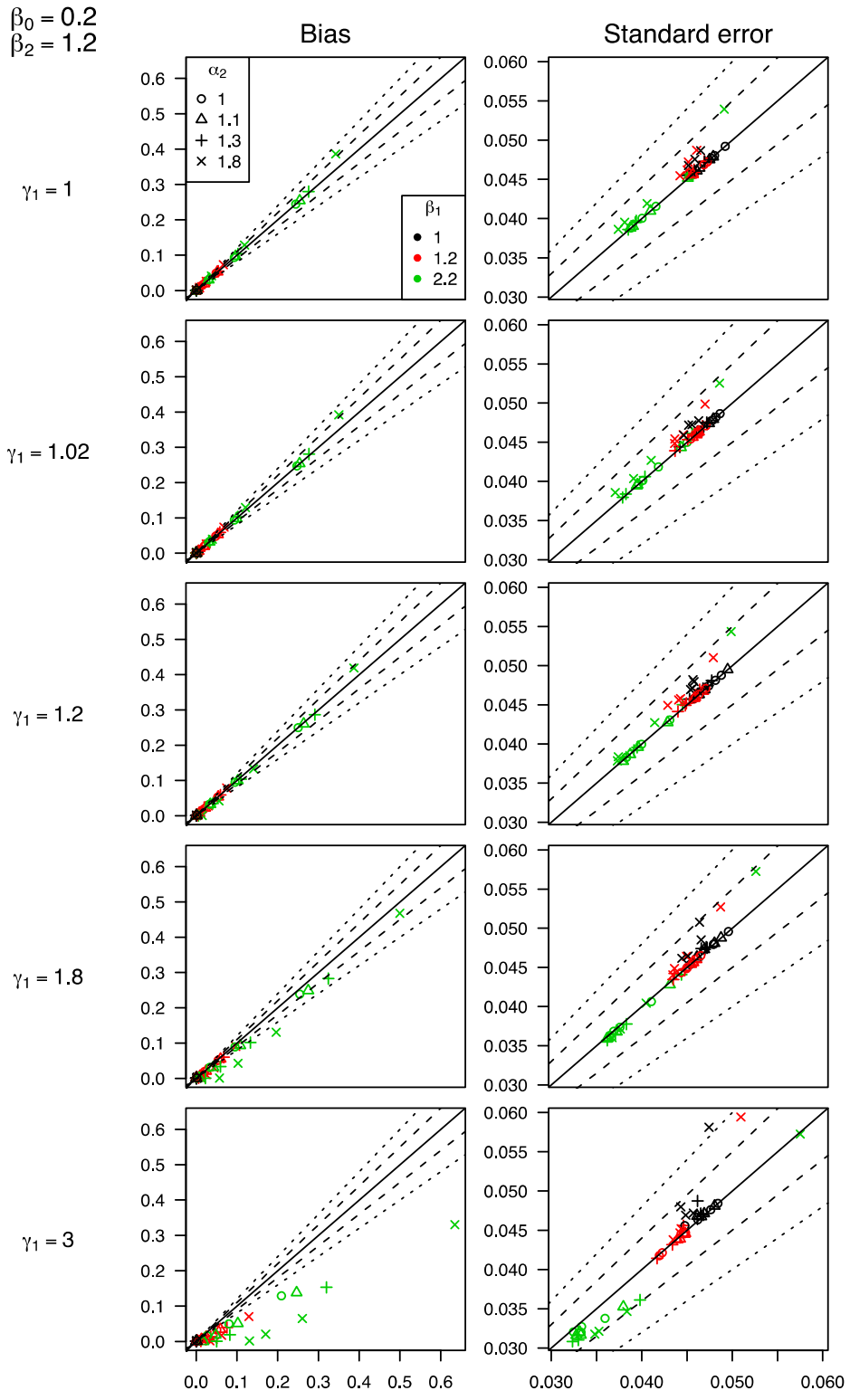
Appendix Figure 6: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.01 and β_2=2.2.



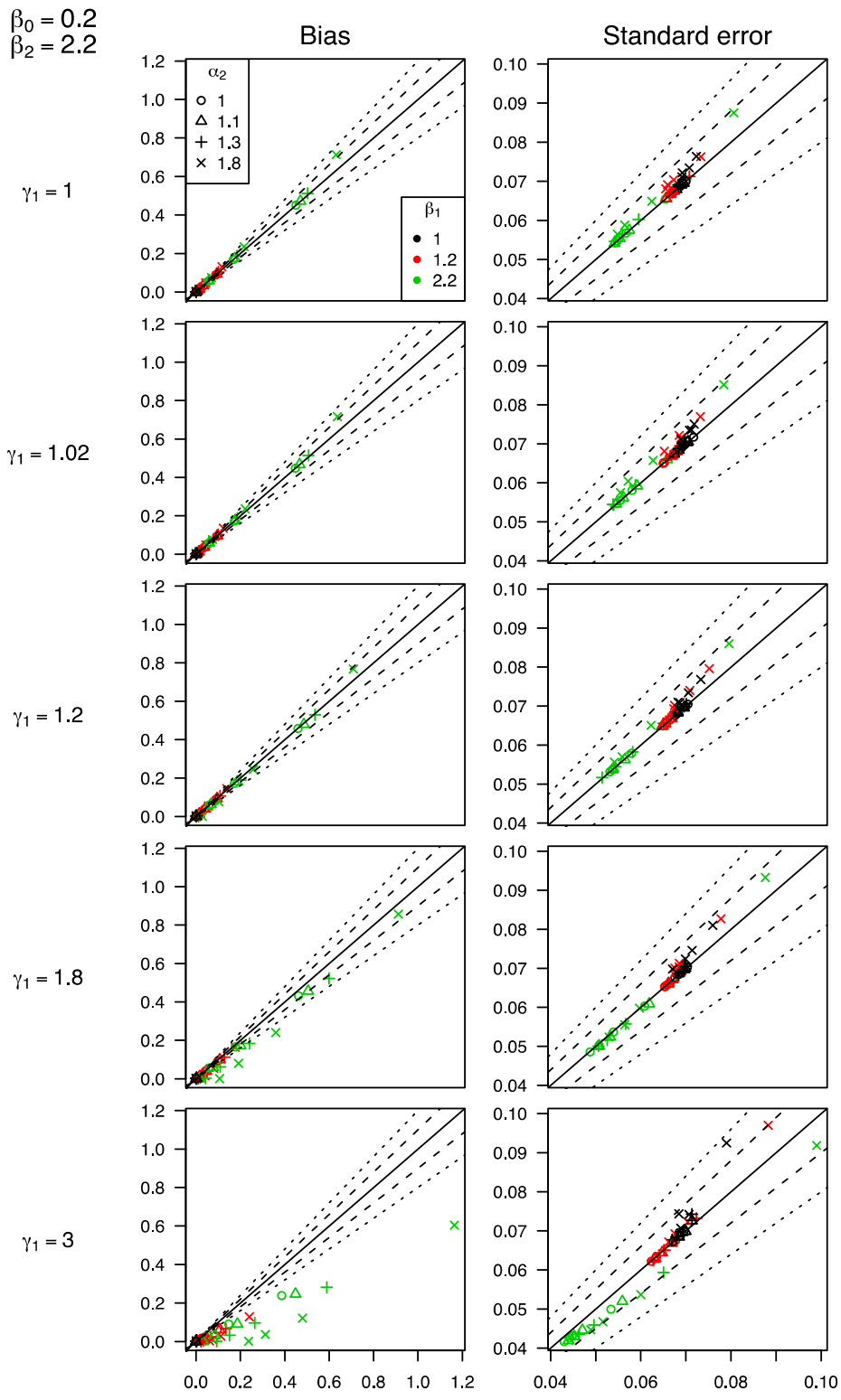
Appendix Figure 7: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.01 and β_2=8.



Appendix Figure 8: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.2 and β_2=1.



Appendix Figure 9: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.2 and β_2=1.2.



Appendix Figure 10: Each point represents one simulation scenario in the multiplicative simulations with β_0=0.2 and β_2=2.2.

D. APPENDIX D: SIMULATION CODE

All data generation, analysis, and plotting was performed in R. In this section, we provide the code used for simulation so that others may reproduce our results. The function addiSims, along with the accompanying functions rd.crude and rd.cond, simulates and analyzes data for one set of simulation parameters in the additive framework. The function multiSims, along with the accompanying functions rr.crude and rr.cond, simulates and analyzes data for one set of simulation parameters in the multiplicative framework.

```
rd.cond <- function(y, x, c) {
    cases1 <- rowsum(y*x, c)
    cases0 <- rowsum(y*(1-x), c)
    n1 <- rowsum(x, c)
    n0 <- rowsum(1-x, c)
    n <- c(sum(1-c), sum(c))
    sum((cases1*n0 - cases0*n1)/n) / sum(n1*n0/n)
}

rd.crude <- function(y, x) {
    n1 <- sum(x)
    n0 <- sum(1-x)
    p1 <- sum(y*x)/n1
    p0 <- sum(y*(1-x))/n0
    p1 - p0
}

addiSims <- function(simpars, nsamp=10000, nsim=2500) {
    results <- matrix(0, nsim, 3)
    dat <- matrix(NA, nsim, 2^3)
    colnames(dat) <- c("z0x0y0", "z0x0y1", "z0x1y0", "z0x1y1",
                       "z1x0y0", "z1x0y1", "z1x1y0", "z1x1y1")

    # simpars should be an R data frame or list with named elements
    # gamma0-beta2
    g0 <- simpars$gamma0
    g1 <- simpars$gamma1
    a0 <- simpars$alpha0
    a1 <- simpars$alpha1
    a2 <- simpars$alpha2
    b0 <- simpars$beta0
    b1 <- simpars$beta1
    b2 <- simpars$beta2

    for(s in 1:nsim){             # make the data
        z <- rbinom(nsamp, 1, .5)
        u <- rbinom(nsamp, 1, g0 + g1*z)
        x <- rbinom(nsamp, 1, a0 + a1*u + a2*z)
        y <- rbinom(nsamp, 1, b0 + b1*u + b2*x)
        dat[s,] <- as.vector(table(y,x,z))
```

```
        # estimates
        results[s,] <- round(c(rd.crude(y, x), #unadjusted association
                               rd.cond(y, x, u), # adjusting for u
                               rd.cond(y, x, z), # adjusting for z
                               ), 6)
    }
    colnames(results) <- c("crude", "truth", "condZ")
    results <- cbind(results, dat)
    results
}

rr.cond <- function(y, x, c) {
    cases1 <- rowsum(y*x, c)
    cases0 <- rowsum(y*(1-x), c)
    n1 <- rowsum(x, c)
    n0 <- rowsum(1-x, c)
    n <- c(sum(1-c), sum(c))
    sum(cases1*n0/n)/sum(cases0*n1/n)
}

rr.crude <- function(y, x) {
    n1 <- sum(x)
    n0 <- sum(1-x)
    p1 <- sum(y*x)/n1
    p0 <- sum(y*(1-x))/n0
    p1/p0
}

multiSims <- function(simpars, nsamp=10000, nsim=2500) {
    results <- matrix(0, nsim, 3)
    dat <- matrix(NA, nsim, 2^3)
    colnames(dat) <- c("z0x0y0", "z0x0y1", "z0x1y0", "z0x1y1",
                       "z1x0y0", "z1x0y1", "z1x1y0", "z1x1y1")

    # simpars should be an R data frame or list with named elements
    # gamma0-beta2
    g0 <- simpars$gamma0
    g1 <- simpars$gamma1
    a0 <- simpars$alpha0
    a1 <- simpars$alpha1
    a2 <- simpars$alpha2
    b0 <- simpars$beta0
    b1 <- simpars$beta1
    b2 <- simpars$beta2

    for(s in 1:nsim){             # make the data
        z <- rbinom(nsamp, 1, .5)
        u <- rbinom(nsamp, 1, g0 * g1^z)
        x <- rbinom(nsamp, 1, a0 * a1^u * a2^z)
        y <- rbinom(nsamp, 1, b0 * b1^u * b2^x)
```

```
        dat[s,] <- as.vector(table(y,x,z))

        # estimates
        results[s,] <- round(c(rr.crude(y, x), #unadjusted association
                               rr.cond(y, x, u), # adjusted for u
                               rr.cond(y, x, z), # adjusted for z
                               ), 6)
    }
    colnames(results) <- c("crude", "truth", "condZ")
    results <- cbind(results, dat)
    results
}
```