Econometric Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition

Martin Huber & Lukáš Lafférs

ECONOMETRIC REVIEWS
2022, VOL. 41, NO. 10, 1141–1163
https://doi.org/10.1080/07474938.2022.2127077

Bounds on direct and indirect effects under treatment/ mediator endogeneity and outcome attrition

Martin Huber[a] and Lukáš Lafférs[b]

[a]Department of Economics, University of Fribourg, Fribourg, Switzerland; [b]Department of Mathematics, Matej Bel University, Banská Bystrica, Slovakia

ABSTRACT
Causal mediation analysis aims at disentangling a treatment effect into an indirect mechanism operating through an intermediate outcome or mediator, as well as the direct effect of the treatment on the outcome of interest. However, the evaluation of direct and indirect effects is frequently complicated by non-ignorable selection into the treatment and/or mediator, even after controlling for observables, as well as sample selection/outcome attrition. We propose a method for bounding direct and indirect effects in the presence of such complications using a method that is based on a sequence of linear programming problems. Considering inverse probability weighting by propensity scores, we compute the weights that would yield identification in the absence of complications and perturb them by an entropy parameter reflecting a specific amount of propensity score misspecification to set-identify the effects of interest. We apply our method to data from the National Longitudinal Survey of Youth 1979 to derive bounds on the explained and unexplained components of a gender wage gap decomposition that is likely prone to non-ignorable mediator selection and outcome attrition.

KEYWORDS
Bounds; causal channels; causal mechanisms; direct effects; indirect effects; mediation analysis; sample selection

JEL CLASSIFICATION
C21

1. Introduction

Mediation analysis aims to decompose a treatment effect into an indirect causal mechanism operating through one or several intermediate variables, so-called mediators, as well as the direct effect, including any mechanisms not operating through the mediators of interest. For instance, early childhood interventions might affect labor market or health outcomes later in life through different mechanisms like the formation of cognitive or non-cognitive skills, see, for instance, Heckman et al. (2013) and Keele et al. (2015). Furthermore, job seeker counseling may influence employment through assignment to training programs or other mechanisms in the counseling process, see Huber et al. (2017). Even with a randomly assigned treatment, direct and indirect effects are generally not identified by naively controlling for mediators, as this likely introduces selection bias, see Robins and Greenland (1992). While much of the earlier work on mediation analysis assumed linear models and/or neglected selection issues, see Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986), more recent contributions discuss more general identification approaches and explicitly consider confounding. See, for instance, Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen et al. (2006), VanderWeele (2009), Imai et al. (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen and Shpitser (2012), and Vansteelandt et al. (2012).

CONTACT Lukáš Lafférs lukas.laffers@gmail.com Department of Mathematics, Matej Bel University, Tajovského 40, SK97401 Banská Bystrica, Slovakia.

In most mediation studies, identification relies on a conditionally exogenous treatment and mediator given observed covariates and rules out non-ignorable outcome attrition or sample selection, i.e., that outcomes are only observed for a nonrandom subpopulation. This issue occurs, for instance, in wage regressions, where wages are only observed for the selective subgroup of employed individuals, see Heckman (1976) and Heckman (1979). For this reason, Huber and Solovyeva (2020a) incorporate outcome attrition into mediation models, assuming conditional treatment and mediator exogeneity and tackling outcome attrition either by observed covariates (missing at random assumption, see, e.g., Little and Rubin, 1987; Rubin, 1976) or by instruments (if attrition is selective in unobservables). In many empirical problems, however, observed covariates might not be rich enough to convincingly control for treatment/mediator endogeneity and attrition bias while instruments that satisfy specific exclusion restrictions w.r.t. attrition (see, for instance, Das et al., 2003; Huber, 2014) might not be available.

This paper provides a method for deriving bounds on direct and indirect effects when the treatment, the mediator and outcome attrition are likely selective even after controlling for observed covariates. Considering identification based on inverse probability weighting based on a combination of propensity scores, we compute the weights that would yield identification in the absence of complications (as provided in Huber and Solovyeva, 2020a) and perturb them by an entropy parameter reflecting misspecification in the various propensity scores. Based on the framing the identification issue as an optimization problem to be solved by linear programming, we set-identify the mean potential outcomes and thus, the direct and indirect effects of interest.

Our contribution is related to further studies that used optimization, and in particular linear programming, to derive bounds on treatment effects under selection problems, see, e.g., Balke and Pearl (1997), Manski (2007), Honoré and Tamer (2006), Molinari (2008), Freyberger and Horowitz (2015), Lafférs and Nedela (2017), Lafférs (2019), among many others. Our paper is also related to the literature on sensitivity analysis in mediation analysis. VanderWeele (2010), for instance, provides a general formula for the bias of direct and indirect effects in the presence of an unobserved mediator-outcome confounder. By considering sensible values for differences in conditional mean outcomes across confounder values and for differences in the conditional mean of the confounder across treatment states, researches may investigate the sensitivity of the effects. Imai et al. (2010) propose a sensitivity check for parametric (both linear and nonlinear) mediation models based on specifying the correlation of unobserved terms in the mediation and outcome equations, assuming that the mediator-outcome confounders are not a function of the treatment. In contrast, Tchetgen and Shpitser (2012) suggest a semiparametric procedure that allows for confounders of the mediator-outcome relation which are affected by the treatment based on specifying and calibrating the so-called selection bias function, which is agnostic about the dimension of unobserved confounders. See VanderWeele and Chiba (2014) and Vansteelandt and VanderWeele (2012) for further selection bias functions.

As an alternative strategy, Albert and Nelson (2011) suggest considering the correlation of counterfactual values of post-treatment variables as sensitivity parameter. Finally, the paper that is the closest to our approach is Hong et al. (2018), which provides a method tailored to weighting estimators under the omission of both pre- and post-treatment confounders. The idea is that such confounders create a discrepancy between the correct weight an observation should obtain and the one actually used. The resulting bias can be represented by the covariance between the weight discrepancy and the outcome conditional on the treatment, which serves as base for conducting sensitivity analyses. Our approach is different to Hong et al. (2018) in that we represent the discrepancy between the correct and observed weights using entropy parameters and, instead of deriving analytical formulas, rely on an optimization routine to obtain bounds on direct and indirect effects. This allows for a separate relaxation of the three main identification assumptions and thus may lead to a better understanding of the non-robustness of the results to violations of the various identification assumptions. We also note that the econometric setup of Huber and

Solovyeva (2020a) underlying our analysis invokes a different set of assumptions than Hong et al. (2018) or any of the other previous methods, such that our approach permits investigating sensitivity also w.r.t. outcome attrition.[1]

We apply our method to data from the National Longitudinal Survey of Youth 1979, a panel study of young individuals in the U.S. aged 14 to 22 years in 1979. The specific sample considered has previously been analyzed by Huber and Solovyeva (2020b) to decompose the gender gap in wages reported in the year 2000 into an indirect (or explained) component due to differences in mediators like education and occupation, as well as a direct (or unexplained) gender difference in wages not attributable to the observed mediators. While Huber and Solovyeva (2020b) investigated the sensitivity of point estimation of explained and unexplained component under different identifying assumptions, our approach permits easing any of the conditional exogeneity assumptions on gender, the mediators, and selection into employment (as wages are only observed for working individuals) to derive bound son the parameters of interest. We find that the omission of confounders of the treatment and the mediators would potentially have the largest impact on the significance of the results. More specifically, the omission of a confounder that has the same predictive power as the first or second most important mediator entering the treatment propensity score would render all the effects insignificant. The results also show that in some specifications the choice of the link function in the estimation of probabilistic weights matter.

The remainder of this paper is organized as follows. Section 2 introduces the variables as well as the direct and indirect effects of interest. Section 3 restates the identifying assumptions of Huber and Solovyeva (2020a), under which the direct and indirect effects are point identified, and introduces the sensitivity analysis based on inverse probability weighting when relaxing these assumptions. Section 4 presents an application to the decomposition of the U.S. gender wage gap using data from the National Longitudinal Survey of Youth 1979. Section 5 concludes.

2. Variables and parameters of interest

Mediation analysis typically aims to disentangle the average treatment effect (ATE) of a binary treatment, denoted by D, on an outcome variable, denoted by Y, into a direct effect and an indirect effect operating through one or several mediators. We denote the latter by M, which is assumed to have bounded support and may be scalar or a vector of variables and contain discrete and/or continuous elements. For defining natural direct and indirect effects, we make use of the potential outcome framework, see, for instance, Rubin (1974), which has been applied to causal mediation analysis, for instance, by Ten Have et al. (2007) and Albert (2008). Let to this end $M(d), Y(d, M(d'))$ denote the potential mediator state as a function of the treatment and potential outcome as a function of the treatment and the potential mediator, respectively, under treatments $d, d' \in \{0, 1\}$. For each subject, only one potential outcome and mediator state, respectively, is observed, because the realized mediator and outcome values are $M = D \cdot M(1) + (1-D) \cdot M(0)$ and $Y = D \cdot Y(1, M(1)) + (1-D) \cdot Y(0, M(0))$.

The ATE, denoted by Δ, is given by the total effect of the treatment operating through the direct or indirect mechanisms:

$$\Delta = E[Y(1, M(1)) - Y(0, M(0))]. \tag{1}$$

The (average) direct effect, denoted by $\theta(d)$, is characterized by the difference in mean potential outcomes under treatment and non-treatment when fixing the mediator at its potential value for $D = d$, which shuts down the indirect mechanism via M.

[1]The studies mentioned and our own investigate the sensitivity of direct and indirect effects to prespecified deviations from the identifying assumptions. Alternatively, one may derive worst case bounds, which are based on the possibly most extreme forms of violations of specific assumptions, which typically implies a rather wide range of admissible effect values. See for instance Kaufman et al. (2005), Cai et al. (2008), Sjölander (2009), and Flores and Flores-Lagunes (2010).

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}. \tag{2}$$

The (average) indirect effect, denoted by $\delta(d)$, is given by the difference in mean potential outcomes when exogenously varying the mediator to take its potential values under treatment and non-treatment, but keeping the treatment fixed at $D = d$ to shut down the direct effect.

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{3}$$

Robins and Greenland (1992) and Robins (2003) referred to these causal parameters as pure/total direct and indirect effects, Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects, and Pearl (2001) as natural direct and indirect effects, which is the denomination followed in the remained of this study.

The ATE is the sum of the natural direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned}\Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1).\end{aligned} \tag{4}$$

This follows from adding and subtracting either $E[Y(0, M(1))]$ or $E[Y(1, M(0))]$ in (4). The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ allows for effect heterogeneity as a function of the treatment state, i.e., the presence of interaction effects between the treatment and the mediator. For instance, the impact of a training (M) on employment (Y) might depend on whether a job seeker has received some form of counseling in the job search process (D). A different way to see this is that the direct effect of counseling (D) may depend on whether the job seeker attends a training (M).

Obviously, effects are not identified without invoking identifying assumptions. First, $Y(1, M(1))$ and $Y(0, M(0))$ are not observed for any subject at the same time, which constitutes the fundamental problem of causal inference. Second, neither $Y(1, M(0))$, nor $Y(0, M(1))$ is observed for any subject. Therefore, point identification of direct and indirect effects requires the treatment and the mediator to be exogenous at least conditional on observables, which appears, however, implausible in many empirical applications. Our sensitivity analysis outlined in Section 3 relaxes such exogeneity conditions at the cost of giving up on point identification. This permits incorporating a vector of observed pretreatment covariates, denoted by X, that may confound the causal relations between D and M, D and Y, and M and Y. It is thus assumed that X is insufficient to control for all sources of selection such that unobserved confounders render point identification of direct and indirect effects impossible, which appears plausible in many empirical contexts.

As a further complication to identification, our framework allows for considering outcome attrition/sample selection, implying that Y is only observed for a nonrandom subpopulation. For instance, when investigating wage outcomes, as in Gronau (1974), the subpopulation of employed individuals for whom wages are observed might be positively selected in terms of unobservables like ability and motivation. As a further example, consider the effect of educational interventions on test scores, with scores being only observed for those participating in the test or reporting the results, see Angrist et al. (2006). We therefore introduce a binary selection indicator S, which indicates whether Y is observed for a specific subject. S is allowed to be a function of D, M, and X, i.e., $S = S(D, M, X)$, but is assumed to neither be affected by nor to affect outcome Y. S is therefore not a mediator, as selection per se does not causally influence the outcome, but might nevertheless create endogeneity bias when outcomes are only observed conditional on $S = 1$. While the outcome Y is not observed for subjects with $S = 0$, we assume D, M, and X to be observed for everyone. This permits identifying the conditional selection probability

$\Pr(S = 1|D, M, X)$, which plays an important role for the sensitivity analysis suggested in the next section.

3. Sensitivity analysis

The starting point for our sensitivity analysis is a set of assumptions provided in Huber and Solovyeva (2020a), which identifies direct and indirect effects by invoking conditional treatment and mediator exogeneity as well outcome attrition related to observed characteristics (known as missing at random assumption). Formally, the assumptions are as follows:

Assumption A1 (conditional independence of the treatment).
(a) $Y(d, m) \perp D|X = x$, (b) $M(d') \perp D|X = x$ for all $d, d' \in \{0, 1\}$ and m, x in the support of M, X.

Assumption A1 rules out unobservables jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on X. In contrast, our sensitivity analysis permits that such unobserved confounders do exist.

Assumption A2 (conditional independence of the mediator).
$Y(d, m) \perp M|D = d', X = x$ for all $d, d' \in \{0, 1\}$ and m, x in the support of M, X.

Assumption A2 rules out unobservables jointly affecting the mediator and the outcome conditional on D and X. This only appears plausible if detailed information on possible confounders of the mediator-outcome relation is available in the data (even in experiments with random treatment assignment) and if post-treatment confounders of M and Y can be plausibly ruled out when controlling for D and X. In contrast, our sensitivity analysis allows for unobserved confounders of the mediator-outcome relation.

Assumption A3 (conditional independence of selection).
$Y \perp S|D = d, M = m, X = x$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

Assumption A3 rules out unobservables jointly affecting selection and the outcome conditional on D, M, X, such that outcomes are missing at random (MAR) in the denomination of Rubin (1976), i.e., outcome attrition is selective w.r.t. observed characteristics only. In contrast, our sensitivity analysis permits outcome attrition to be selective w.r.t. unobservables.

Assumption A4 (common support).
(a) $\Pr(D = d|M = m, X = x) > 0$ and (b) $\Pr(S = 1|D = d, M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

Assumption A4 consists of two common support restrictions. The first requires the conditional probability to receive a specific treatment given M, X, henceforth referred to as propensity score, to be larger than zero for either treatment state. This also implies that $\Pr(D = d|X = x) > 0$ and (by Bayes' theorem) that $\Pr(M = m|D = d, X = x) > 0$, or in the case of M being continuous, that the conditional density of M given D, X is larger than zero. Therefore, M must not be deterministic in D given X, as otherwise identification fails due to the lack of comparable units in terms of the mediator across treatment states. The second common support restriction requires that for any combination of D, M, X, the probability to be observed is larger than zero. Otherwise, the outcome is not observed for some specific combinations of these variables. Our sensitivity relies on the same set of common support assumptions, in order to make treated and non-treated subjects with observed and non-observed outcomes comparable in terms of observed characteristics.

As outlined in Theorem 1 of Huber and Solovyeva (2020a), Assumptions A1 to A4 permit identifying the mean potential outcomes based on inverse probability weighting (IPW) by

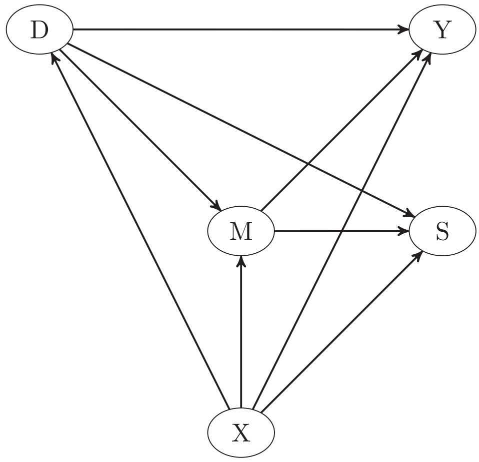

Figure 1. Causal paths under conditional exogeneity and missing at random given pretreatment covariates.

$$E[Y(1,M(1))] = E\left[Y \cdot D \cdot S \cdot \frac{1}{\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)}\right],$$
$$E[Y(0,M(0))] = E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{1-\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)}\right],$$
$$E[Y(1,M(0))] = E\left[Y \cdot D \cdot S \cdot \frac{1}{1-\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)} \cdot \left(\frac{1}{\Pr(D=1|M,X)} - 1\right)\right],$$
$$E[Y(0,M(1))] = E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)} \cdot \left(\frac{1}{1-\Pr(D=1|M,X)} - 1\right)\right]. \quad (5)$$

The direct and indirect effects of interest are obtained as differences between two out of the four mean potential outcomes. For notational ease, we henceforth denote the various propensity scores in (5) by

$$p^{A1} = \Pr(D=1|X), \quad p^{A2} = \Pr(D=1|M,X), \quad p^{A3} = \Pr(S=1|D,M,X). \quad (6)$$

This denomination is motivated by the fact that, e.g., under A3, there are no confounders that jointly affect S and D conditional on (M, X), S and M conditional on (D, X) or S and X conditional on (M, D). So under A3, p^{A3} is the correct probability to be used for weighting in order to obtain mean potential outcomes. Conversely, if A3 does not hold, e.g., some important confounder is missing, then the correct weight differs from p^{A3}.

Figure 1 illustrates our mediation framework with outcome attrition based on a directed acyclic graph, in which the arrows represent causal effects. It is worth noting that each of D, M, S, X, and Y might be causally affected by further, unobserved variables not displayed in Fig. 1. As long as such unobservables do not jointly affect D and Y, D and M, M and Y, or S and Y conditional on X, Assumptions A1 to A3 hold. In many empirical problems, however, some or all of A1 to A3 appear difficult defend. While, for instance, A1 holds by design in randomized experiments, it may appear less plausible in observational studies, in particular if the set of observed control variables is limited. Assumption A2 might seem unlikely in any identification design, in particular if there is a substantial time lag between D and M which makes mediator-outcome confounding more likely even when conditioning on pretreatment covariates X. We therefore consider various relaxations of Assumptions A1 to A3.

Our approach modifies the IPW weights of the expressions in (5) to investigate sensitivity and is thus related to Hong et al. (2018), who were the first to propose robustness checks in the context of IPW. However, our approach uses a different measure of discrepancies between the correct and observed weights than Hong et al. (2018), whose sensitivity check is based on the covariance between the weight discrepancy and the outcome conditional on the treatment. Also, our approach does not entail analytical formulae and thus relies on optimization routines. Even though this increases the computational burden, an advantage is that we are able to separately consider relaxations of the various identifying assumptions and thus gain insights on the sources of the potential non-robustness of the effects.

To see the intuition of our approach, suppose for the moment that there exist a scalar of a vector of unobserved confounders, denoted by U, that makes Assumption A3 fail. In this case, p^{A3} is no longer the appropriate propensity score for identifying the mean potential outcomes. Our sensitivity analysis is based on specifying the magnitude by which p^{A3} may deviate from the appropriate (but unidentified) propensity score that includes U as conditioning variable. More formally, we consider the following entropy measure defined as the absolute difference between p^{A3} and the appropriate propensity score for identification, $q^{A3} = \Pr(S=1|D,M,X,U)$:

$$|q^{A3} - p^{A3}| \leq \epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}. \tag{7}$$

This definition restricts the absolute error in the propensity score p^{A3} due to omitting confounders U to a multiple ϵ^{A3} of the standard deviation of a random variable with a binomial distribution corresponding to that of p^{A3}. This definition also satisfies a symmetry property, so that the relaxation of the identifying assumption leads to the same set of values for p^{A3} and for $(1-p^{A3})$. The crucial task is to sensibly choose the value of ϵ^{A3}, e.g., based on the richness of X, which determines the likely importance of omitted confounders U. In an analogous way, $q^{A1} = \Pr(D=1|X,U)$, $q^{A2} = \Pr(D=1|M,X,U)$ as well as ϵ^{A1} and ϵ^{A2} are to be defined. Suppose there were no unobserved confounders and that Assumptions A1, A2, A3 were satisfied. That would correspond to the situation with $\epsilon^{A1} = \epsilon^{A3} = \epsilon^{A3} = 0$. The greater is the particular entropy parameter, the larger is the permitted importance of unobserved confounders in the specific assumption.

There are many different ways of modeling the relaxation of identifying assumptions. Our approach is closely related to the total variation distance metric between two distributions, with the distinction that we scale differences in our probabilities based on the standard deviation. More specifically, we bound the maximum error coming from confounding due to erroneously considering p^{A3} as the correct propensity rather than the true one q^{A3}. A straightforward bounding measure is a baseline distance parameter (e.g., ϵ^{A3}) that is multiplied with a probability's standard deviation, which takes into account differences in the variances across propensity scores that are closer or less close to zero or one. Our choice is also driven by practical considerations, as the optimization problem inherent in our sensitivity analysis is computationally attractive and can be analyzed using readily available software.

Our sensitivity analysis provides bounds on estimates of the mean potential outcomes in (5) for violations of A1, A2, and A3, subject to the restrictions implied by ϵ^{A1}, ϵ^{A2}, and ϵ^{A3} in (7) as well as specific scaling constraints. The bounds on mean potential outcomes then translate into bounds on natural direct and indirect effects. When considering, for instance, the mean potential outcome $E[Y(1,M(1))]$, note that we may express the latter under the Assumptions A1, A3 and A4 as

$$E[Y(1,M(1))] = E\left[\frac{Y\cdot D\cdot S}{\Pr(D=1|X)\cdot\Pr(S=1|D,M,X)}\right] = E\left[\frac{Y\cdot D\cdot S}{p^{A1}\cdot p^{A3}}\right], \tag{8}$$

see Eq. (6) in Huber and Solovyeva (2020a). However, in the case that Assumptions A1 or A3 are not satisfied due to confounding by an unobserved variable U, the mean potential outcome corresponds to

$$E[Y(1,M(1))] = E\left[\frac{Y \cdot D \cdot S}{\Pr(D=1|X,U) \cdot \Pr(S=1|D,M,X,U)}\right] = E\left[\frac{Y \cdot D \cdot S}{q^{A1} \cdot q^{A3}}\right]. \tag{9}$$

Our method aims at determining a range (or set) of possible values of $E[Y(1,M(1))]$ in (9), given that the discrepancy of the propensities under confounding (q^{A1}, q^{A3}) and no confounding (p^{A1}, p^{A3}) can be bounded by specific inequality constraints like (7). The population optimization problem is the following:

$$\min_{q^{A1},q^{A3}}/\max\; E[Y(1,M(1))] = E\left[\frac{Y \cdot D \cdot S}{q^{A1} \cdot q^{A3}}\right].$$

$$s.t.$$

$$|q^{A1} - p^{A1}| \le \epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A3} - p^{A3}| \le \epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

In Appendix A, we present the optimization problems for deriving bounds on all four mean potential outcomes $E[Y(1,M(1))]$, $E[Y(1,M(0))]$, $E[Y(0,M(1))]$, and $E[Y(0,M(0))]$ in the population under violations of the identifying assumptions.

We subsequently discuss the estimation procedure. Assume that we have available an i.i.d. sample of $(Y_i, D_i, M_i, X_i, S_i)$ for n subjects, where $i \in \{1,2,\ldots,n\}$ indexes a specific observation. Under Assumptions A1, A3, and A4, this quantity can be estimated by

$$\hat{E}[Y(1,M(1))] = \frac{1}{c} \cdot \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \frac{1}{\hat{p}_i^{A1}} \cdot \frac{1}{\hat{p}_i^{A3}},$$

where $\hat{p}_i^{A1}, \hat{p}_i^{A3}$ denote estimates of p^{A1}, p^{A3} for observation i, which we obtain in our application presented below by logit regression, and c denotes a normalizing constant, $c = \sum_{i=1}^{n} \frac{D_i}{\hat{p}_i^{A1}} \cdot \frac{S_i}{\hat{p}_i^{A3}}$.[2] In the presence of confounders U and a failure of assumptions A1 and/or A3, estimating $\hat{E}[Y(1,M(1))]$ based on $\hat{p}_i^{A1}, \hat{p}_i^{A3}$ is generally inconsistent. To estimate the bounds on the mean potential outcome under such violations, the unknown population parameters p^{A1} and p^{A3} are replaced by their estimates $\hat{p}_i^{A1}$ and $\hat{p}_i^{A3}$ in (7) to estimate the entropy measures $|q^{A1} - p^{A1}|$ and $|q^{A3} - p^{A3}|$, respectively.

Estimating the bounds on $\hat{E}[Y(1,M(1))]$ corresponds to solving the following optimization problem in the sample:

$$\min_{q^{A1},q^{A3}}/\max \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \frac{1}{q_i^{A1}} \cdot \frac{1}{q_i^{A3}} \Big/ \sum_{i=1}^{n} \frac{D_i}{q_i^{A1}} \frac{S_i}{q_i^{A3}} \tag{10}$$

$$s.t.$$

$$\forall i: \; |q_i^{A1} - \hat{p}_i^{A1}| \le \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}, \quad |q_i^{A3} - \hat{p}_i^{A3}| \le \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}, \tag{11}$$

$$q_i^{A1} \in [0,1], \quad q_i^{A3} \in [0,1]. \tag{12}$$

The normalization restriction is now implicit in (10), while the expressions in (11) relax the identifying assumptions and (12) ensures that q_i are proper probabilities. We note that as only those observations i with $D_i = 1$ and $S_i = 1$ enter the calculations, we added a superscript 1 to the entropy parameters ϵ (for $D_i = 1$). This implies that one might pick different parameters for different treatment groups, if, e.g., justified by contextual knowledge.

[2]Note that Assumption A2 is not required for the identification of $E[Y(d,M(d))]$ for $d \in \{0,1\}$, but for $E[Y(d,M(1-d))]$.

The optimization problem (10) may be transformed into a computationally more convenient form. Let to this end denote $z_i = \frac{1}{q_i^{A1} q_i^{A3}}$. Then, an alternative representation is

$$\begin{aligned}
&\min_z/\max \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot z_i \Big/ \sum_{i=1}^{n} D_i \cdot S_i \cdot z_i \\
&\quad s.t. \\
&\forall i : z_i \leq 1 \Big/ \left(\left(\hat{p}_i^{A1} - \epsilon^{A1,1} \sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} - \epsilon^{A3,1} \sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\
&\quad z_i \geq 1 \Big/ \left(\left(\hat{p}_i^{A1} + \epsilon^{A1,1} \sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} + \epsilon^{A3,1} \sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\
&\quad z_i \geq 1.
\end{aligned} \tag{13}$$

This simplification is feasible because there is no interplay between the restrictions on q^{A1} and q^{A3}, which enter the objective function only as a product. The resulting optimization problem is a linear-fractional program that can be translated with to the following linear program (Charnes and Cooper, 1962) by introducing an additional auxiliary variable t:

$$\begin{aligned}
&\min_{\omega, t}/\max \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i \\
&\quad s.t. \\
&\forall i : \omega_i \leq t \Big/ \left(\left(\hat{p}_i^{A1} - \epsilon^{A1,1} \sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} - \epsilon^{A3,1} \sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\
&\quad \omega_i \geq t \Big/ \left(\left(\hat{p}_i^{A1} + \epsilon^{A1,1} \sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} + \epsilon^{A3,1} \sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\
&\sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i = 1, \quad \omega_i \geq 0, \quad t \geq 0.
\end{aligned}$$

An important feature of our approach based on the optimization is that we impose no structure on the dependence of potentially omitted confounders and our outcome variable and exhaust all possibilities for the weights. This may in general lead to wider bounds and thus more prudent inference.

Bounds on $\hat{E}[Y(0, M(0))]$ (i.e., the estimate of $E[Y(0, M(0))]$) can be constructed in an analogous manner by using observations with $S_i = 1$, $D_i = 0$ and searching through $(1 - q_i^{A1})$ instead. For bounds on $\hat{E}[Y(1, M(0))]$ and $\hat{E}[Y(0, M(1))]$, one has to optimize w.r.t. q_i^{A1}, q_i^{A2} and q_i^{A3}. These optimization problems can again be simplified in a similar manner and all linear programming formulations are stated in Appendix A. After deriving upper and lower bounds on the mean potential outcomes, we may construct bounds on the effects of interest in the following way:

$$\begin{aligned}
\hat{\Delta}_{LB} &= \hat{E}[Y(1, M(1))]_{LB} - \hat{E}[Y(0, M(0))]_{UB}, \\
\hat{\Delta}_{UB} &= \hat{E}[Y(1, M(1))]_{UB} - \hat{E}[Y(0, M(0))]_{LB}, \\
\hat{\theta}(1)_{LB} &= \hat{E}[Y(1, M(1))]_{LB} - \hat{E}[Y(0, M(1))]_{UB}, \\
\hat{\theta}(1)_{UB} &= \hat{E}[Y(1, M(1))]_{UB} - \hat{E}[Y(0, M(1))]_{LB}, \\
\hat{\theta}(0)_{LB} &= \hat{E}[Y(1, M(0))]_{LB} - \hat{E}[Y(0, M(0))]_{UB}, \\
\hat{\theta}(0)_{UB} &= \hat{E}[Y(1, M(0))]_{UB} - \hat{E}[Y(0, M(0))]_{LB}, \\
\hat{\delta}(1)_{LB} &= \hat{E}[Y(1, M(1))]_{LB} - \hat{E}[Y(1, M(0))]_{UB}, \\
\hat{\delta}(1)_{UB} &= \hat{E}[Y(1, M(1))]_{UB} - \hat{E}[Y(1, M(0))]_{LB}, \\
\hat{\delta}(0)_{LB} &= \hat{E}[Y(0, M(1))]_{LB} - \hat{E}[Y(0, M(0))]_{UB}, \\
\hat{\delta}(0)_{UB} &= \hat{E}[Y(0, M(1))]_{UB} - \hat{E}[Y(0, M(0))]_{LB},
\end{aligned}$$

where subscripts LB, UB stand for the lower and upper bounds of the respective estimated mean potential outcome and "^" implies that any of the causal effects are estimates rather than population parameters. The bounds on Δ, $\theta(1)$, and $\theta(0)$ are sharp, because the observations that are used for calculating the two mean potential outcomes upon which the respective effect is defined are distinct. Consider, for instance, the lower bound on $\theta(1)$. In order to estimate $E[Y(1, M(1))]_{LB}$, we use observations with $D_i = 1$, while for $E[Y(0, M(1))]_{UB}$ we only use observations with $D_i = 0$. In contrast, the bounds for $\delta(1)$ and $\delta(0)$ are not necessarily sharp. As an example, consider $\delta(1)$. In order to estimate $E[Y(1, M(1))]_{LB}$ and $E[Y(1, M(0))]_{UB}$, we use observations with $D_i = 1$ and the weights that entail $E[Y(1, M(1))]_{LB}$ and $E[Y(1, M(0))]_{UB}$ are not necessary the same.[3]

An important question yet to be discussed is how to set the entropy parameters, which represent changes in the propensity scores due to omitting confounders (e.g., $\epsilon^{A1,1}$ and $\epsilon^{A3,1}$), in a meaningful way. Investigating the importance of observed confounders X may provide some guidance for finding sensible values for the entropy parameters. Consider, for instance, a logistic regression for estimating p_i^{A3} :

$$S_i \sim Bern(p_i^{A3}),$$
$$\log\left(\frac{p_i^{A3}}{1 - p_i^{A3}}\right) = \alpha_0 + \alpha_D D_i + \alpha_M^T M_i + \alpha_X^T X_i,$$

where $\hat{p}_i^{A3}$ is obtained by maximum likelihood estimation. Now consider removing the most important predictor (of S) in X and re-estimating the propensity score, denoted as $\hat{p}_{i,X1}^{A3}$, where $X1$ are the remaining covariates (without the most important predictor).

For each observation, one can then compute the entropy parameter when including and excluding the most important predictor in X, $\epsilon_{i,X1}^{A3} = \frac{|\hat{p}_{i,X1}^{A3} - \hat{p}_i^{A3}|}{\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$. One may ultimately pick the entropy parameter as average of $\epsilon_{i,X1}^{A3}$ in the subpopulation with $D_i = 1$ and $S_i = 1$:

$$\epsilon_{X1}^{A3,1} = \sum_{i=1}^{n} \frac{D_i \cdot S_i \cdot \epsilon_{i,X1}^{A3}}{\sum_{i=1}^{n} D_i \cdot S_i}.$$

This corresponds to the average change induced by omitting the most important predictor from X, which is used as a proxy for the importance of unobserved confounders U. There are different ways of measuring the importance of a predictor in a regression and natural choice seems to be the change in the deviance. The latter is the log-likelihood ratio statistic for testing the difference in the model fit when including and excluding a specific predictor, which is asymptotically chi-squared distributed. Similarly ϵ_{Xj}^{A3} and ϵ_{Mj}^{A3} would denote the average change in estimated probabilities that would omitting the j-th most important (measured the by deviance) from X and M, respectively.[4]

4. Application

We apply our method to data from the National Longitudinal Survey of Youth 1979 (NLSY79), a panel survey of young individuals who were aged 14 to 22 years in 1979, to decompose the gender wage gap in the year 2000 when respondents were 35 to 43 years old.[5] We use exactly the

[3]It is in principle possible to compute sharp bounds even for $\delta(1)$ and $\delta(0)$ at the cost of solving a more complex optimization problem. This issue is discussed in a greater detail in Appendix B.

[4]Another approach for setting the entropy parameter is to consider the change in estimated probabilities induced by a change of the link function, e.g. by picking the probit instead of logit function. Formally, $\epsilon_{i,probit}^{A3,1} = \frac{|\hat{p}_{i,probit}^{A3} - \hat{p}_i^{A3}|}{\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$, where $\hat{p}_{i,probit}^{A3}$ and $\hat{p}_i^{A3}$ correspond to the estimated probabilites under a probit and logit model, respectively. One may thus pick the entropy parameter as average $\epsilon_{probit}^{A3,1} = \sum_{i=1}^{n} \frac{D_i \cdot S_i \cdot \epsilon_{i,probit}^{A3}}{\sum_{i=1}^{n} D_i \cdot S_i}$.

[5]The NLSY79 data consists of three independent samples: a cross-sectional sample (6,111 subjects, or 48%) representing the non-institutionalized civilian youth; a supplemental sample (42%) oversampling civilian Hispanic, black, and economically disadvantaged nonblack/non-Hispanic young people; and a military sample (10%) comprised of youth serving in the military as of September 30, 1978 (Bureau of Labor Statistics, U.S. Department of Labor, 2001).

same sample definition as in Huber and Solovyeva (2020b), who consider five wage decomposition techniques with distinct identifying assumptions to investigate the sensitivity of point estimators of the indirect effect (or explained component) due to gender differences in mediators like education or occupation as well as the direct gender effect (or unexplained component). However, the consistency of these decomposition techniques relies on specific conditional exogeneity or instrumental variable assumptions w.r.t. to gender, the mediators, and the observability of wages, which are likely violated in practice. See also Huber (2015) for a discussion of identification issues with in wage decompositions based on the causal mediation framework. In contrast, the approach suggested in this paper permits investigating the robustness of the direct and indirect effects of gender on wage under violations of conditional exogeneity.

The NLSY79 includes a rich set of individual characteristics, including socio-economic variables likes education, occupation, work experience, and a range of further labor market-relevant information. Our evaluation sample consists of 6,658 individuals (3,162 men and 3,496 women), after excluding 1,351 observations from the total NLSY79 sample in 2000 due to various data issues outlined in Huber and Solovyeva (2020b).[6] Treatment D is a binary indicator for gender, taking the value zero for females and one for males. Outcome Y corresponds to the log average hourly wage in the past calendar year reported in 2000. However, the wage outcome under full-time employment is not observed for everyone, as a non-negligible share (in particular among females) are in minor employment or not employed at all. This likely introduces sample selection issues, see Heckman (1979). We therefore define the selection indicator S to be one for individuals who indicated to have worked at least 1,000 hours in the past calendar year, as it is the case for 87% of males and 70% of females.

The vector of mediators M includes individual variables reported in or constructed with reference to 1998 such that they arguably reflect decisions taken after birth and prior to the measurement of the outcome (i.e., on the causal path between D and Y): marital status, years in marriage, the region and the number of years residing in that region, a dummy for living in an urban area (SMSA) and the number of years living in an urban area, education, dummies for the year of first employment, number of jobs ever had, tenure with the current employer (in weeks), industry and the number of years working in that industry, occupation and the number of years working in that occupation, whether employed in 1998 and total years of employment, a dummy for full-time employment and the share of full-time employment from 1994–98, total weeks of employment, the number of weeks unemployed and the number of weeks out of the labor force, and whether health problems prevented employment along with the history of health problems.

In the propensity scores, we control for a set of observed covariates X arguably mostly determined at or prior to birth, namely race, religion, year of birth, birth order, parental place of birth (in the U.S. or abroad), and parental education. However, unobserved confounders causing non-ignorable selection into the treatment, mediators, and/or employment decision S even after controlling for X likely exist. For instance, risk preferences, attitudes toward competition and negotiations, and other socio-psychological factors, see, e.g., Bertrand (2011) and Azmat and Petrongolo (2014), are not available in our data. Such variables might possibly confound the mediator-outcome relationship. As a second example, selection into employment might be driven by innate ability or motivation, which likely also affect wages. Due to such endogeneity concerns, one should be cautious about deriving causal claims (e.g., about the amount of labor market discrimination) and policy recommendations based on wage gap decompositions, see the discussion in Huber and Solovyeva (2020b). In this context, our method is useful for assessing the sensitivity of the results to violations of some or all exogeneity conditions required for a causal interpretation of wage decompositions.

[6]For instance, we excluded 502 persons reporting to work 1,000 hours or more in the past calendar year, but whose average hourly wages in the past calendar year were either missing or equal to zero. Furthermore, we dropped 54 working individuals with average hourly wages of less than $1 in the past calendar year. We also excluded 608 observations with missing values in the mediators.

Table 1. Covariates and mediators with the highest predictive power.

Assumption A1		$P(D=1\|X)$
Most important X	1st	Mothers educ. missing
	2nd	Mothers educ. high school graduate
	3rd	Religion missing
Assumption A2		$P(D=1\|M,X)$
Most important M	1st	Farmer or laborer
	2nd	Industry: Professional services
	3rd	Clerical occupation
Most important X	1st	White
	2nd	Fathers educ. college/more
	3rd	Mothers educ. missing
Assumption A3		$P(S=1\|D,M,X)$
Most important M	1st	Employed full time
	2nd	Employment status: employed
	3rd	Operator (machines, transport)
Most important X	1st	Fathers educ. college/more
	2nd	Mothers educ. some college
	3rd	Protestant

Note: Most important predictors in different propensity score estimations measured by the change in deviance.

Table D1 in Appendix D provides descriptive statistics for the key variables of our analysis, namely means, mean differences across gender, and the respective p-values based on two-sample t-tests. The p-values suggest that females an males differ importantly in a range of variables like labor market experience, education, industry, and occupation. Our sensitivity analysis is based on assuming that omitted confounders in the various propensity score specifications behave similar to the first, second, or third important predictors among the pretreatment covariates or mediators that enter the respective specification. For this reason, Table 1 shows the three most important covariates and mediators in the different propensity score models presumably prone to confounding, where importance refers to the change in deviance as discussed at the end of Section 3.

Tables 2–5 present the estimated effect bounds on $\theta(1)$, $\theta(0)$, $\delta(1)$, and $\delta(0)$, respectively, in squared brackets, when basing the entropy parameter on the respective first, second, or third most important predictors.[7] 95% confidence intervals are also reported in parentheses, based on subsampling bootstrap without replacement with 1000 replications and a subsample size of $\lfloor n^{0.7} \rfloor$.[8] We observe that the direct effects remain statistically significant at the 5% level under most relaxations and that potentially missing confounders would have the biggest impact via Assumption A2. Previous employment has considerable explanatory power for later labor market performance, as reflected in the relatively wide bounds in the column for the 2nd most important missing M (the variable *Industry: professional services*) in Table 2, where the lower 95% confidence bound goes below zero. Table 3 displays the estimated bounds for the natural direct effect under $d=0$, when mediators are set to their potential values for females and the indirect mechanisms are shut down. These are generally larger than the direct effects under $d=1$ (males), but have qualitatively similar patterns.

Concerning the natural indirect effects reported in Tables 4 and 5, the confidence interval on the effect estimate of 0.053 includes the zero under most relaxations for males ($d=1$), while the lower confidence bound remains above zero under most relaxations for females ($d=0$). We also observe that confidence intervals are not necessarily symmetric around the estimated bounds and

[7]We note that in the decomposition literature, it is frequently the male wages that are considered as reference (or 'fair') wages. This suggests considering $\theta(0)$ and $\delta(1)$ as unexplained and explained components, respectively. See Sloczynski (2013) for an in-depth discussion of reference group choice in the potential outcome framework.

[8]We applied subsampling to the upper and lower bounds separately similar to Lafférs and Nedela (2017) or Demuynck (2015). A computationally more expensive stepdown procedure of Romano and Shaikh (2010) may be used to control for the asymptotic coverage of the whole identified set. Neither the subsample size nor the number of subsamples affected our results in an important way. For completeness, we describe the procedure in detail in Appendix C.

Table 2. Bounds on Natural direct effect for $d=1$ using relaxation parameters that correspond to average deviation that missing the 1st, 2nd and the 3rd most important regressor from X or M would induce.

	Missing X			
Importance	1st	2nd	3rd	Probit
A1	[0.182 0.219]	[0.167 0.235]	[0.194 0.208]	[0.200 0.201]
	(0.096 0.314)	(0.082 0.328)	(0.108 0.303)	(0.115 0.296)
A2	[0.162 0.240]	[0.164 0.238]	[0.179 0.222]	[0.175 0.227]
	(0.080 0.331)	(0.082 0.329)	(0.096 0.314)	(0.092 0.319)
A3	[0.189 0.212]	[0.191 0.210]	[0.192 0.210]	[0.190 0.212]
	(0.104 0.308)	(0.106 0.306)	(0.107 0.306)	(0.104 0.308)
A1 + A2	[0.143 0.258]	[0.130 0.272]	[0.172 0.229]	[0.175 0.227]
	(0.063 0.349)	(0.050 0.362)	(0.090 0.321)	(0.092 0.319)
A2 + A3	[0.150 0.251]	[0.154 0.248]	[0.170 0.231]	[0.164 0.238]
	(0.071 0.339)	(0.074 0.336)	(0.089 0.321)	(0.083 0.327)
A1 + A3	[0.170 0.231]	[0.157 0.244]	[0.184 0.217]	[0.189 0.212]
	(0.086 0.327)	(0.073 0.339)	(0.100 0.313)	(0.104 0.308)
A1 + A2 + A3	[0.131 0.270]	[0.120 0.281]	[0.163 0.238]	[0.164 0.238]
	(0.053 0.358)	(0.042 0.370)	(0.082 0.328)	(0.082 0.327)
	Missing M			
Importance	1st	2nd	3rd	
A2	[0.106 0.240]	[0.056 0.238]	[0.072 0.222]	
	(0.030 0.331)	(−0.018 0.329)	(−0.002 0.314)	
A3	[0.159 0.212]	[0.170 0.210]	[0.179 0.210]	
	(0.077 0.308)	(0.086 0.306)	(0.095 0.306)	
A2 + A3	[0.066 0.251]	[0.026 0.248]	[0.051 0.231]	
	(−0.008 0.339)	(−0.046 0.336)	(−0.022 0.321)	

Bounds on the natural direct effect under $d=1$ (point estimator is 0.201 with 95% CI = (0.115,0.296)).
Last column corresponds to change due to different choice of the link function—probit instead of logit. Different lines stand for relaxations of different identifying assumptions: "A1," "A2," "A3": denote Assumption 1, 2, 3, respectively, and "A1 + A2" denotes simultaneous relaxation of both Assumptions 1 and 2, and similarly for other lines.

Table 3. Bounds on Natural direct effect for $d=0$ using relaxation parameters that correspond to average deviation that missing the 1st, 2nd and the 3rd most important regressor from X or M would induce.

	Missing X			
Importance	1st	2nd	3rd	Probit
A1	[0.307 0.344]	[0.293 0.358]	[0.319 0.332]	[0.325 0.326]
	(0.188 0.461)	(0.174 0.476)	(0.199 0.450)	(0.205 0.444)
A2	[0.282 0.367]	[0.285 0.364]	[0.303 0.347]	[0.299 0.351]
	(0.162 0.476)	(0.165 0.473)	(0.183 0.459)	(0.180 0.462)
A3	[0.305 0.347]	[0.308 0.343]	[0.309 0.342]	[0.306 0.345]
	(0.190 0.467)	(0.193 0.463)	(0.194 0.462)	(0.191 0.465)
A1 + A2	[0.264 0.385]	[0.252 0.397]	[0.296 0.354]	[0.299 0.351]
	(0.144 0.494)	(0.133 0.506)	(0.177 0.465)	(0.180 0.462)
A2 + A3	[0.262 0.389]	[0.268 0.382]	[0.287 0.364]	[0.280 0.371]
	(0.148 0.500)	(0.154 0.493)	(0.173 0.479)	(0.167 0.484)
A1 + A3	[0.287 0.366]	[0.276 0.376]	[0.302 0.349]	[0.306 0.346]
	(0.173 0.485)	(0.161 0.495)	(0.187 0.469)	(0.191 0.466)
A1 + A2 + A3	[0.244 0.407]	[0.235 0.415]	[0.280 0.371]	[0.280 0.371]
	(0.131 0.518)	(0.122 0.526)	(0.167 0.485)	(0.167 0.485)
	Missing M			
Importance	1st	2nd	3rd	
A2	[0.118 0.367]	[0.139 0.364]	[0.202 0.347]	
	(−0.012 0.476)	(0.013 0.473)	(0.080 0.459)	
A3	[0.253 0.347]	[0.271 0.343]	[0.289 0.342]	
	(0.152 0.467)	(0.169 0.463)	(0.180 0.462)	
A2 + A3	[0.044 0.389]	[0.085 0.382]	[0.165 0.364]	
	(−0.069 0.500)	(−0.028 0.493)	(0.052 0.479)	

Bounds on the natural direct effect under $d=0$ (point estimator is 0.325 with 95% CI = (0.205,0.444)).
Last column corresponds to change due to different choice of the link function—probit instead of logit. Different lines stand for relaxations of different identifying assumptions: "A1," "A2," "A3": denote Assumption 1, 2, 3, respectively, and "A1 + A2" denotes simultaneous relaxation of both Assumptions 1 and 2, and similarly for other lines.

Table 4. Bounds on Natural indirect effect for $d = 1$ using relaxation parameters that correspond to average deviation that missing the 1st, 2nd and the 3rd most important regressor from X or M would induce.

	Missing X			
Importance	1st	2nd	3rd	Probit
A1	[0.034 0.071]	[0.020 0.085]	[0.045 0.061]	[0.053 0.053]
	(−0.069 0.154)	(−0.082 0.168)	(−0.059 0.144)	(−0.052 0.136)
A2	[0.011 0.096]	[0.014 0.093]	[0.031 0.075]	[0.027 0.079]
	(−0.080 0.178)	(−0.078 0.176)	(−0.066 0.158)	(−0.069 0.162)
A3	[0.043 0.062]	[0.045 0.061]	[0.045 0.060]	[0.042 0.063]
	(−0.060 0.141)	(−0.058 0.139)	(−0.058 0.139)	(−0.061 0.142)
A1 + A2	[−0.007 0.114]	[−0.019 0.126]	[0.023 0.083]	[0.027 0.079]
	(−0.097 0.196)	(−0.109 0.207)	(−0.074 0.166)	(−0.069 0.162)
A2 + A3	[0.002 0.106]	[0.006 0.101]	[0.023 0.083]	[0.017 0.090]
	(−0.089 0.182)	(−0.086 0.179)	(−0.072 0.161)	(−0.078 0.167)
A1 + A3	[0.025 0.080]	[0.012 0.093]	[0.037 0.068]	[0.042 0.064]
	(−0.077 0.159)	(−0.089 0.171)	(−0.065 0.147)	(−0.061 0.142)
A1 + A2 + A3	[−0.016 0.124]	[−0.027 0.134]	[0.015 0.091]	[0.016 0.090]
	(−0.106 0.199)	(−0.116 0.209)	(−0.080 0.169)	(−0.078 0.168)
	Missing M			
Importance	1st	2nd	3rd	
A2	[−0.101 0.096]	[−0.090 0.093]	[−0.053 0.075]	
	(−0.172 0.178)	(−0.163 0.176)	(−0.129 0.158)	
A3	[0.027 0.062]	[0.030 0.061]	[0.031 0.060]	
	(−0.076 0.141)	(−0.074 0.139)	(−0.073 0.139)	
A2 + A3	[−0.126 0.106]	[−0.113 0.101]	[−0.075 0.083]	
	(−0.199 0.182)	(−0.186 0.179)	(−0.150 0.161)	

Bounds on the natural indirect effect under $d = 1$ (point estimator is 0.053 with 95% CI = (−0.052,0.136)).
Last column corresponds to change due to different choice of the link function—probit instead of logit. Different lines stand for relaxations of different identifying assumptions: "A1," "A2," "A3": denote Assumption 1, 2, 3, respectively, and "A1 + A2" denotes simultaneous relaxation of both Assumptions 1 and 2, and similarly for other lines.

Table 5. Bounds on Natural indirect effect for $d = 0$ using relaxation parameters that correspond to average deviation that missing the 1st, 2nd and the 3rd most important regressor from X or M would induce.

	Missing X			
Importance	1st	2nd	3rd	Probit
A1	[0.159 0.196]	[0.143 0.212]	[0.172 0.183]	[0.177 0.178]
	(0.069 0.283)	(0.054 0.299)	(0.082 0.270)	(0.087 0.264)
A2	[0.138 0.216]	[0.140 0.214]	[0.156 0.199]	[0.152 0.203]
	(0.056 0.296)	(0.058 0.295)	(0.072 0.282)	(0.068 0.284)
A3	[0.155 0.201]	[0.159 0.197]	[0.160 0.196]	[0.158 0.198]
	(0.069 0.289)	(0.073 0.284)	(0.074 0.283)	(0.072 0.285)
A1 + A2	[0.120 0.235]	[0.106 0.249]	[0.150 0.205]	[0.151 0.204]
	(0.038 0.315)	(0.024 0.329)	(0.066 0.288)	(0.068 0.285)
A2 + A3	[0.116 0.240]	[0.122 0.234]	[0.138 0.217]	[0.132 0.224]
	(0.035 0.324)	(0.040 0.317)	(0.054 0.301)	(0.050 0.308)
A1 + A3	[0.136 0.220]	[0.125 0.231]	[0.154 0.202]	[0.158 0.198]
	(0.051 0.307)	(0.037 0.318)	(0.068 0.288)	(0.072 0.285)
A1 + A2 + A3	[0.097 0.259]	[0.087 0.268]	[0.133 0.223]	[0.132 0.224]
	(0.016 0.343)	(0.006 0.352)	(0.049 0.307)	(0.049 0.308)
	Missing M			
Importance	1st	2nd	3rd	
A2	[0.082 0.216]	[0.027 0.214]	[0.045 0.199]	
	(0.003 0.296)	(−0.054 0.295)	(−0.035 0.282)	
A3	[0.077 0.201]	[0.111 0.197]	[0.141 0.196]	
	(0.002 0.289)	(0.032 0.284)	(0.057 0.283)	
A2 + A3	[−0.021 0.240]	[−0.040 0.234]	[0.007 0.217]	
	(−0.090 0.324)	(−0.113 0.317)	(−0.066 0.301)	

Bounds on the natural indirect effect under $d = 0$ (point estimator is 0.177 with 95% CI = (0.088,0.264)).
Last column corresponds to change due to different choice of the link function—probit instead of logit. Different lines stand for relaxations of different identifying assumptions: "A1," "A2," "A3": denote Assumption 1, 2, 3, respectively, and "A1 + A2" denotes simultaneous relaxation of both Assumptions 1 and 2, and similarly for other lines.

that the omission of the variable with the most predictive power (measured by the change in deviance) does not necessarily lead to the widest bounds. The latter is due to the fact that, e.g., the strongest predictor of the treatment is not necessarily the strongest confounder, i.e., the predictor's association with the outcome is sufficiently weaker than that of other treatment predictors. The choice of the link function does not seem overwhelmingly crucial for the bounds. In fact, if we permit the difference between the correct and estimated weights to be of a similar magnitude as the difference due to using a probit instead of a logit specification, then the bounds are similar to those obtained under a confounder with similar predictive power as the third most important covariate in X. We also observe that concerning Assumption A1, the choice of the link function is close to irrelevant.

5. Conclusion

This paper proposed a sensitivity check for estimating natural direct and indirect effects in the presence of treatment and mediator endogeneity as well as selective attrition or missingness in outcomes. To this end, we considered identification based on inverse probability weighting using treatment and selection propensity scores and perturbed the respective propensity scores by an entropy parameter reflecting a specific amount of misspecification to set-identify the effects of interest. We demonstrated that this approach can be framed as a linear programming problem and discussed sensible choices of the entropy parameters based on the predictive power of the most important predictors in the propensity scores. Finally, we applied our method to data from the National Longitudinal Survey of Youth 1979 to derive bounds on the explained and unexplained components of a gender wage gap decomposition that is likely prone to non-ignorable mediator selection and sample selection in terms of the observability of the wage outcomes.

Appendices

Appendix A. Deriving bounds based on linear programming

A.1. Optimization problems for bounds on mean potential outcomes

For $i = 1, \dots, n$, $K \in \{1, 2, 3\}$ and for a particular relaxation type $R \in \{X1, X2, X3, M1, M2, M3, probit\}$ we denote

- $\omega_i^{AK} = 1/q_i^{AK}$
- $\bar{\omega}_i^{AK} = 1/(1 - q_i^{AK})$,
- $\epsilon_{i,R}^{AK} = \frac{|\hat{p}_{i,R}^{AK} - \hat{p}_i^{AK}|}{\sqrt{\hat{p}_i^{AK}(1-\hat{p}_i^{AK})}}$
- $\epsilon_R^{AK,1} = \sum_{i=1}^{n} \frac{D_i \cdot S_i \cdot \epsilon_{i,R}^{AK}}{\sum_{i=1}^{n} D_i \cdot S_i}$
- $\epsilon_R^{AK,0} = \sum_{i=1}^{n} \frac{(1-D_i) \cdot S_i \cdot \epsilon_{i,R}^{AK}}{\sum_{i=1}^{n} (1-D_i) \cdot S_i}$

A.2. Bounds on $\hat{E}[Y(1, M(1))]$

A.2.1. Population optimization problem

$$\min_{q^{A1}, q^{A3}}/\max E[Y(1, M(1))] = E\left[\frac{Y \cdot D \cdot S}{\Pr(D = 1|X, U) \cdot \Pr(S = 1|D, M, X, U)}\right] = E\left[\frac{Y \cdot D \cdot S}{q^{A1} \cdot q^{A3}}\right].$$

$$s.t.$$

$$|q^{A1} - p^{A1}| \leq \epsilon^{A1} \sqrt{p^{A1}(1 - p^{A1})},$$

$$|q^{A3} - p^{A3}| \leq \epsilon^{A3} \sqrt{p^{A3}(1 - p^{A3})}.$$

A.2.2. Optimization problem in the sample

$$\min_{\omega,t}/\max\sum_{i=1}^{n} Y_i\cdot D_i\cdot S_i\cdot\omega_i$$

$$s.t.$$

$$\forall i:\omega_i\leq t/\left(\left(\hat{p}_i^{A1}-\epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)\cdot\left(\hat{p}_i^{A3}-\epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)\right),$$

$$\omega_i\geq t/\left(\left(\hat{p}_i^{A1}+\epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)\cdot\left(\hat{p}_i^{A3}+\epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)\right),$$

$$\sum_{i=1}^{n} D_i\cdot S_i\cdot\omega_i=1,\quad \omega_i\geq 0,\quad t\geq 0,$$

where $\omega_i=\frac{1}{q_i^{A1}\cdot q_i^{A3}}$.

A.3. Bounds on $\hat{E}[Y(0,M(0))]$

A.3.1. Population optimization problem

$$\min_{q^{A1},q^{A3}}/\max\ E[Y(0,M(0))]=E\left[Y\cdot(1-D)\cdot S\cdot\frac{1}{1-\Pr(D=1|X,U)}\cdot\frac{1}{\Pr(S=1|D,M,X,U)}\right]=E\left[\frac{Y\cdot(1-D)\cdot S}{(1-q^{A1})\cdot q^{A3}}\right].$$

$$s.t.$$

$$|q^{A1}-p^{A1}|\leq\epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A3}-p^{A3}|\leq\epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

A.3.2. Optimization problem in the sample

$$\min_{\omega,t}/\max\sum_{i=1}^{n} Y_i\cdot(1-D_i)\cdot S_i\cdot\omega_i$$

$$s.t.$$

$$\forall i:\omega_i\leq t/\left(\left(1-\hat{p}_i^{A1}-\epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)\cdot\left(\hat{p}_i^{A3}-\epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)\right),$$

$$\omega_i\geq t/\left(\left(1-\hat{p}_i^{A1}+\epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)\cdot\left(\hat{p}_i^{A3}+\epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)\right),$$

$$\sum_{i=1}^{n}(1-D_i)\cdot S_i\cdot\omega_i=1,\quad \omega_i\geq 0,\quad t\geq 0,$$

where $\omega_i=\frac{1}{(1-q_i^{A1})\cdot q_i^{A3}}$.

A.4. Bounds on $\hat{E}[Y(1,M(0))]$

A.4.1. Population optimization problem

$$E[Y(1,M(0))]=E\left[Y\cdot D\cdot S\cdot\frac{1}{1-\Pr(D=1|X)}\cdot\left(\frac{1}{\Pr(D=1|M,X)}-1\right)\cdot\frac{1}{\Pr(S=1|D,M,X)}\right].$$

$$=E\left[Y\cdot D\cdot S\cdot\frac{1}{1-q^{A1}}\cdot\left(\frac{1}{q^{A2}}-1\right)\cdot\frac{1}{q^{A3}}\right]$$

$$s.t.$$

$$|q^{A1}-p^{A1}|\leq\epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A2}-p^{A2}|\leq\epsilon^{A2}\sqrt{p^{A2}(1-p^{A2})},$$
$$|q^{A3}-p^{A3}|\leq\epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

A.4.2. Optimization problem in the sample

$$\min_{\omega, t}/\max\sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i$$

$$s.t.$$

$$\forall i: \omega_i \le t \cdot \frac{1}{1-\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{\hat{p}_i^{A2} - \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}},$$

$$\omega_i \ge t \cdot \frac{1}{1-\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{\hat{p}_i^{A2} + \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}},$$

$$\sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i = 1, \quad \omega_i \ge 0, \quad t \ge 0,$$

where $\omega_i = \frac{1}{1-q_i^{A1}} \cdot \left(\frac{1}{q_i^{A2}} - 1\right) \cdot \frac{1}{q_i^{A3}}$.

A.5. Bounds on $\hat{E}[Y(0, M(1))]$

A.5.1. Population optimization problem

$$E[Y(0,M(1))] = E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{\Pr(D=1|X)} \cdot \left(\frac{1}{1-\Pr(D=1|M,X)} - 1\right) \cdot \frac{1}{\Pr(S=1|D,M,X)}\right].$$

$$= E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{q^{A1}} \cdot \left(\frac{1}{1-q^{A2}} - 1\right) \cdot \frac{1}{q^{A3}}\right]$$

$$s.t.$$

$$|q^{A1} - p^{A1}| \le \epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A2} - p^{A2}| \le \epsilon^{A2}\sqrt{p^{A2}(1-p^{A2})},$$
$$|q^{A3} - p^{A3}| \le \epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

A.5.2. Optimization problem in the sample

$$\min_{\omega, t}/\max\sum_{i=1}^{n} Y_i \cdot (1-D_i) \cdot S_i \cdot \omega_i$$

$$s.t.$$

$$\forall i: \omega_i \le t \cdot \frac{1}{\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{1-\hat{p}_i^{A2} - \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$$

$$\omega_i \ge t \cdot \frac{1}{\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{1-\hat{p}_i^{A2} + \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$$

$$\sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i = 1, \quad \omega_i \ge 0, \quad t \ge 0.$$

where $\omega_i = \frac{1}{q_i^{A1}} \cdot \left(\frac{1}{1-q_i^{A2}} - 1\right) \cdot \frac{1}{q_i^{A3}}$.

Appendix B. Discussion on the sharpness of the bounds

This section discusses in greater detail why our bounds on direct effects are sharp, while the bounds on the indirect effects are not necessarily sharp. Concerning the direct effect, we consider the bounds on $\theta(1)$, i.e., the natural direct effect under treatment, and note that analogous arguments hold for $\theta(0)$. The optimization problem for obtaining the bounds in the population is given below.

B.1. Population optimization problem

$$\min_{q^{A1},q^{A2},q^{A3}}/\max \quad \theta(1) = E[Y(1,M(1))] - E[Y(0,M(1))] =$$

$$E\left[Y \cdot D \cdot S \cdot \underbrace{\frac{1}{q^{A1}} \cdot \frac{1}{q^{A3}}}_{\omega^{11}}\right] - E\left[Y \cdot (1-D) \cdot S \cdot \underbrace{\frac{1}{q^{A1}} \cdot \left(\frac{1}{1-q^{A2}} - 1\right) \cdot \frac{1}{q^{A3}}}_{\omega^{01}}\right].$$

$$s.t.$$

$$|q^{A1} - p^{A1}| \leq \epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A2} - p^{A2}| \leq \epsilon^{A2}\sqrt{p^{A2}(1-p^{A2})},$$
$$|q^{A3} - p^{A3}| \leq \epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

We note that ω^{11} and ω^{01} are associated via q^{A1} and q^{A3}. However, ω^{11} is used in the optimization only if $D=1$, while ω^{01} is exclusively considered under $D=0$. Therefore, their interdependence may be ignored for estimation purposes and the optimization problem in the sample outlined below may be split into two distinct optimization problems, as described in subsections A.2 and A.5.

B.2. Optimization problem in the sample

$$\min_{\omega,t}/\max \quad \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i^{11} - \sum_{i=1}^{n} Y_i \cdot (1-D_i) \cdot S_i \cdot \omega_i^{01}$$

$$s.t.$$

$$\forall i: \omega_i^{11} \leq t^{11} \frac{1}{\left(\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)} \cdot \frac{1}{\left(\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)},$$

$$\omega_i^{11} \geq t^{11} \frac{1}{\left(\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)} \cdot \frac{1}{\left(\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)},$$

$$\forall i: \omega_i^{01} \leq t^{01} \cdot \frac{1}{\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{1-\hat{p}_i^{A2} - \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$$

$$\omega_i^{01} \geq t^{01} \cdot \frac{1}{\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{1-\hat{p}_i^{A2} + \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$$

$$\sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i^{11} = 1, \quad \omega_i^{11} \geq 0, \quad t^{11} \geq 0,$$

$$\sum_{i=1}^{n} (1-D_i) \cdot S_i \cdot \omega_i^{01} = 1, \quad \omega_i^{01} \geq 0, \quad t^{01} \geq 0.$$

Let us now consider the bounds on the natural indirect effect under treatment, i.e., $\delta(1)$, and notice that analogous arguments also hold for $\delta(0)$. The optimization problem for obtaining the bounds in the population is given below.

B.3. Population optimization problem

$$\min_{q^{A1},q^{A2},q^{A3}}/\max \quad \delta(1) = E[Y(1,M(1))] - E[Y(1,M(0))] =$$

$$E[Y \cdot D \cdot S \cdot \underbrace{\frac{1}{q^{A1}} \cdot \frac{1}{q^{A3}}}_{\omega^{11}}] - E[Y \cdot D \cdot S \cdot \underbrace{\frac{1}{1-q^{A1}} \cdot (\frac{1}{q^{A2}} - 1) \cdot \frac{1}{q^{A3}}}_{\omega^{10}}].$$

$$s.t.$$

$$|q^{A1} - p^{A1}| \leq \epsilon^{A1}\sqrt{p^{A1}(1-p^{A1})},$$
$$|q^{A2} - p^{A2}| \leq \epsilon^{A2}\sqrt{p^{A2}(1-p^{A2})},$$
$$|q^{A3} - p^{A3}| \leq \epsilon^{A3}\sqrt{p^{A3}(1-p^{A3})}.$$

We see that both ω^{11} and ω^{10} are simultaneously used for the optimization when $D=1$ and therefore, their dependence cannot be ignored. The following optimization problem for estimating the bounds in the sample ignores this dependence and thus only provides *valid*, but generally not sharp bounds on $\delta(1)$.

B.4. Optimization problem in the sample

$$\min_{\omega,t}/\max \quad \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i^{11} - \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i^{10}$$

$$s.t.$$

$$\forall i: \omega_i^{11} \leq t^{11} \frac{1}{\left(\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)} \cdot \frac{1}{\left(\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)},$$

$$\omega_i^{11} \geq t^{11} \frac{1}{\left(\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}\right)} \cdot \frac{1}{\left(\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}\right)},$$

$$\forall i: \omega_i^{10} \leq t^{10} \cdot \frac{1}{1-\hat{p}_i^{A1} - \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{\hat{p}_i^{A2} - \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} - \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}},$$

$$\omega_i^{10} \geq t^{10} \cdot \frac{1}{1-\hat{p}_i^{A1} + \epsilon^{A1,1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}} \cdot \left(\frac{1}{\hat{p}_i^{A2} + \epsilon_R^{A2,1}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})}} - 1\right) \cdot \frac{1}{\hat{p}_i^{A3} + \epsilon^{A3,1}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}},$$

$$\sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i^{11} = 1, \quad \omega_i^{11} \geq 0, \quad t^{11} \geq 0,$$

$$\sum_{i=1}^{n} (1-D_i) \cdot S_i \cdot \omega_i^{10} = 1, \quad \omega_i^{10} \geq 0, \quad t^{10} \geq 0.$$

The dependence between ω^{11} and ω^{10} cannot be captured via linear restrictions and therefore, the optimization problem can no longer be formulated as a linear program.

In principle, we could avoid any transformations and solve this optimization problem directly:

$$\min_{q^{A1},q^{A2},q^{A3}}/\max \quad \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \left(\frac{1}{q_i^{A1}} \cdot \frac{1}{q_i^{A3}} \Big/ \sum_{i=1}^{n} \frac{D_i}{q_i^{A1}} \frac{S_i}{q_i^{A3}} - \frac{1}{1-q_i^{A1}} \cdot \left(\frac{1}{q_i^{A2}} - 1\right) \cdot \frac{1}{q_i^{A3}} \Big/ \sum_{i=1}^{n} \frac{D_i}{1-q_i^{A1}} \cdot \left(\frac{1}{q_i^{A2}} - 1\right) \cdot \frac{S_i}{q_i^{A3}}\right)$$

$$s.t.$$

$$\forall i: |q_i^{A1} - \hat{p}_i^{A1}| \leq \epsilon^{A1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})},$$
$$|q_i^{A2} - \hat{p}_i^{A2}| \leq \epsilon^{A2}\sqrt{\hat{p}_i^{A2}(1-\hat{p}_i^{A2})},$$
$$|q_i^{A3} - \hat{p}_i^{A3}| \leq \epsilon^{A3}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}.$$
$$q_i^{A1} \in [0,1], \; q_i^{A2} \in [0,1], \quad q_i^{A3} \in [0,1].$$

While these constraints are linear, the objective function is, however, highly non-linear such that this optimization is in general computationally expensive.

Appendix C. Confidence intervals based on subsampling

This section describes the construction of the 95% confidence intervals based on subsampling. Consider a parameter of interest, denoted by θ, which may represent a natural direct or indirect effect under $d=1$ or $d=0$, or the total effect. Under the relaxation of the identifying assumptions we have that $\theta \in [\underline{\theta}, \bar{\theta}]$, where $\underline{\theta}$ and $\bar{\theta}$ denote the upper and lower bounds in the population. The estimates of the bounds in the sample are denoted by $[\hat{\underline{\theta}}, \hat{\bar{\theta}}]$. We apply subsampling separately to $\underline{\theta}$ and $\bar{\theta}$. Our $(1-\alpha)$—confidence interval for $[\underline{\theta}, \bar{\theta}]$ consists of the lower limit of the confidence interval for $\underline{\theta}$ and the upper limit of the confidence interval for $\bar{\theta}$. The subsampling procedure below is justified in Politis et al. (1999; see Chapter 2) under mild regularity conditions.

1. Choose the subsample size b_n such that $b_n \to \infty$ and $b_n/n \to 0$ is satisfied. Our choice in the application is $b_n = \lfloor n^{0.7} \rfloor$.
2. Draw b_n observations without replacement from the data and repeat this step B times. We set the number of subsamples to $B = 1000$.
3. Estimate the upper and lower bounds in every subsample, which yields $\hat{\underline{\theta}}^1, \hat{\underline{\theta}}^2, \ldots, \hat{\underline{\theta}}^B$ and $\hat{\bar{\theta}}^1, \hat{\bar{\theta}}^2, \ldots, \hat{\bar{\theta}}^B$.
4. Calculate $L_n(t) = \frac{1}{B}\sum_{i=1}^{B} I(\sqrt{b_n}(\hat{\underline{\theta}}^i - \hat{\underline{\theta}}) \le t)$ and $U_n(t) = \frac{1}{B}\sum_{i=1}^{B} I(\sqrt{b_n}(\hat{\bar{\theta}}^i - \hat{\bar{\theta}}) \le t)$ and denote by $\hat{l}_{1-\alpha/2}$ the $(1-\alpha/2)$ quantile of L_n and by $\hat{u}_{\alpha/2}$ the $\alpha/2$ quantile of U_n.
5. Compute the confidence interval by

$$\left[\hat{\underline{\theta}} - \frac{\hat{l}_{1-\alpha/2}}{\sqrt{n}}, \hat{\bar{\theta}} - \frac{\hat{u}_{\alpha/2}}{\sqrt{n}}\right].$$

Appendix D. Descriptive statistics of the application

Table D1. Summary statistics and mean differences by gender.

Variables	Male($D=1$)	Female($D=0$)	Difference	p-value
Outcome Y (non-logged, refers to selected population with S = 1)				
Hourly wage	19.370	14.164	5.206	0.000
Mediators M (refer to 1998 unless otherwise is stated)				
Married	0.566	0.568	−0.002	0.882
Years married total since 1979	6.430	7.537	−1.107	0.000
Northeastern region	0.153	0.155	−0.002	0.857
North Central region	0.242	0.237	0.005	0.602
West region	0.206	0.195	0.011	0.244
South region (ref.)	0.399	0.414	−0.015	0.205
Years lived in current region since 1979	14.839	15.246	−0.407	0.000
Resides in SMSA	0.811	0.816	−0.005	0.584
Years lived in SMSA since 1979	13.488	14.201	−0.713	0.000
Less than high school (ref.)	0.129	0.101	0.028	0.000
High school graduate	0.459	0.416	0.043	0.000
Some college	0.208	0.271	−0.063	0.000
College or more	0.204	0.213	−0.009	0.413
First job before 1975	0.065	0.046	0.019	0.001
First job in 1976–79	0.115	0.128	−0.013	0.083
First job after 1979 (ref.)	0.821	0.825	−0.004	0.623
Numer of jobs ever had	10.555	9.239	1.316	0.000
Tenure with current employer (wks.)	276.056	212.662	63.394	0.000
Industry: Primary sector	0.227	0.078	0.149	0.000
Industry: Manufacturing (ref.)	0.140	0.053	0.087	0.000
Industry: Transport	0.115	0.048	0.067	0.000
Industry: Trade	0.134	0.142	−0.008	0.322
Industry: Finance	0.040	0.064	−0.024	0.000
Industry: Services (business, personnel, and entertain.)	0.121	0.124	−0.003	0.768
Industry: Professional services	0.113	0.297	−0.184	0.000
Industry: Public administration	0.054	0.052	0.002	0.751
Years worked in current industry since 1982	3.555	2.622	0.933	0.000
Manager	0.234	0.258	−0.024	0.022
Technical occupation (ref.)	0.039	0.038	0.001	0.907
Occupation in sales	0.067	0.082	−0.015	0.021
Clerical occupation	0.056	0.212	−0.156	0.000
Occupation in service	0.102	0.163	−0.061	0.000
Farmer or laborer	0.276	0.042	0.234	0.000

(*continued*)

Table D1. Continued.

Variables	Male($D = 1$)	Female($D = 0$)	Difference	p-value
Operator (machines, transport)	0.170	0.063	0.107	0.000
Years worked in current occupation since 1982	2.180	1.727	0.453	0.000
Employment status: employed	0.877	0.748	0.129	0.000
Number of years employed status since 1979	13.204	11.271	1.933	0.000
Employed full time	0.846	0.599	0.247	0.000
Share of full-time employment 1994–98	0.896	0.658	0.238	0.000
Total number of weeks worked since 1979	661.794	560.408	101.386	0.000
Total number of weeks unemployed since 1979	62.343	49.744	12.599	0.000
Total number of weeks out of labor force since 1979	146.118	265.276	−119.158	0.000
Bad health prevents from working	0.045	0.055	−0.010	0.071
Years not working due to bad health since 1979	0.326	0.557	−0.231	0.000
Pretreatment covariates X				
Hispanic (ref.)	0.193	0.186	0.007	0.488
Black	0.287	0.297	−0.010	0.413
White	0.520	0.517	0.003	0.840
Born in the U.S.	0.935	0.939	−0.004	0.544
No religion	0.045	0.034	0.011	0.031
Protestant	0.501	0.500	0.001	0.957
Catholic (ref.)	0.352	0.352	0.000	0.967
Other religion	0.096	0.112	−0.016	0.036
Mother born in U.S.	0.884	0.896	−0.012	0.102
Mother's educ. <high school (ref.)	0.376	0.421	−0.045	0.000
Mother's educ. high school graduate	0.393	0.369	0.024	0.048
Mother's educ. some college	0.094	0.091	0.003	0.616
Mother's educ. college/more	0.076	0.071	0.005	0.411
Father born in U.S.	0.878	0.884	−0.006	0.410
Father's educ. <high school (ref.)	0.351	0.366	−0.015	0.201
Father's educ. high school graduate	0.291	0.297	−0.006	0.560
Father's educ. some college	0.087	0.076	0.011	0.105
Father's educ. college/more	0.131	0.117	0.014	0.085
Order of birth	3.195	3.259	−0.064	0.256
Age in 1979	17.501	17.611	−0.110	0.047
Selection indicator S				
Worked 1,000 hrs or more past year	0.867	0.696	0.171	0.000
Number of observations	3,162	3,496	.	.

Acknowledgments

Authors gratefully acknowledge the helpful comments of two anonymous referees and the editor Esfandiar Maasoumi.

Declaration of interest statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the Slovak Research and Development Agency under the Grant APVV-17-0329 and under the Grant VEGA-1/0692/20.

References

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine* 27(8):1282–1304. doi: 10.1002/sim.3016

Albert, J. M., Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics* 67(3):1028–1038. doi:10.1111/j.1541-0420.2010.01547.x

Angrist, J., Bettinger, E., Kremer, M. (2006). Long-Term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review* 96(3):847–862. doi:10.1257/aer.96.3.847

Azmat, G., Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics* 30:32–40. doi:10.1016/j.labeco.2014.06.005

Balke, A., Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439):1171–1176. doi:10.1080/01621459.1997.10474074

Baron, R. M., Kenny, D. A. (1986). The Moderator-Mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51(6): 1173–1182. doi:10.1037/0022-3514.51.6.1173

Bertrand, M. (2011). New perspectives on gender, In: Ashenfelter, O., and Card, D., eds., *Handbook of Labor Economics*, Amsterdam, Netherlands: Elsevier, pp. 1543–1590.

Bureau of Labor Statistics, U.S. Department of Labor. (2001). *National Longitudinal Survey of Youth 1979 Cohort, 1979-2000 (Rounds 1–19)*. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

Cai, Z., Kuroki, M., Pearl, J., Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* 64(3):695–701. doi:10.1111/j.1541-0420.2007.00949.x

Charnes, A., Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly* 9(3–4):181–186. doi:10.1002/nav.3800090303

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* 13(3):261–281. doi:10.2307/2527916

Das, M., Newey, W. K., Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies* 70(1):33–58. doi:10.1111/1467-937X.00236

Demuynck, T. (2015). Bounding average treatment effects: a linear programming approach. *Economics Letters* 137: 75–77. doi:10.1016/j.econlet.2015.09.042

Flores, C. A., Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. Iza Dp No. 4237.

Flores, C. A., Flores-Lagunes, A. (2010). *Nonparametric partial identification of causal net and mechanism average treatment effects*. Mimeo, University of Florida.

Freyberger, J., Horowitz, J. L. (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics* 189(1):41–53. doi:10.1016/j.jeconom.2015.06.020

Gronau, R. (1974). Wage comparisons-a selectivity bias. *Journal of Political Economy* 82(6):1119–1143. doi:10.1086/260267

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1):153–161. doi:10.2307/1912352

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

Heckman, J., Pinto, R., Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *The American Economic review* 103(6):2052–2086. doi:10.1257/aer.103.6.2052

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects, in *Proceedings of the American Statistical Association, Biometrics Section*, pp. 2401–2415. Alexandria, VA: American Statistical Association.

Hong, G., Qin, X., Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics* 43(1):32–56. doi:10.3102/1076998617749561

Honoré, B., Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica* 74(3): 611–629. doi:10.1111/j.1468-0262.2006.00676.x

Huber, M. (2014). Treatment evaluation in the presence of sample selection. *Econometric Reviews* 33(8):869–905. doi:10.1080/07474938.2013.806197

Huber, M. (2015). Causal pitfalls in the decomposition of wage gaps. *Journal of Business & Economic Statistics* 33(2):179–191. doi:10.1080/07350015.2014.937437

Huber, M., Lechner, M., Mellace, G. (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics* 99(1):180–183. doi:10.1162/REST_a_00632

Huber, M., Solovyeva, A. (2020a). Direct and indirect effects under sample selection and outcome attrition. *Econometrics* 8(4):44. doi:10.3390/econometrics8040044

Huber, M., Solovyeva, A. (2020b). On the sensitivity of wage gap decompositions. *Journal of Labor Research* 41(1–2):1–33. doi:10.1007/s12122-020-09302-7

Imai, K., Keele, L., Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 25(1):51–71. doi:10.1214/10-STS321

Imai, K., Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis* 21(2):141–171. doi:10.1093/pan/mps040

Judd, C. M., Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* 5(5):602–619. doi:10.1177/0193841X8100500502

Kaufman, S., Kaufman, J. S., MacLehose, R. F., Greenland, S., Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine* 24(11):1683–1702. doi:10.1002/sim.2057

Keele, L., Tingley, D., Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management* 34(4):937–963. doi:10.1002/pam.21853

Lafférs, L. (2019). Bounding average treatment effects using linear programming. *Empirical Economics* 57(3): 727–767. doi:10.1007/s00181-018-1474-z

Lafférs, L., Nedela, R. (2017). Sensitivity of the bounds on the ATE in the presence of sample selection. *Economics Letters* 158:84–87. doi:10.1016/j.econlet.2017.06.039

Little, R., Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review* 48(4):1393–1410. doi:10.1111/j.1468-2354.2007.00467.x

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics* 144(1):81–117. doi:10.1016/j.jeconom.2007.12.003

Pearl, J. (2001). Direct and indirect effects, In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., Sinisi, S. E., van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology* 17: 276–284.

Politis, D. N., Romano, J. P., Wolf, M. (1999). *Subsampling*. Berlin/Heidelberg, Germany: Springer Science & Business Media.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects, In Green, P., Hjort, N., and Richardson, S. *In Highly Structured Stochastic Systems*, pp. 70–81, Oxford: Oxford University Press.

Robins, J. M., Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology (Cambridge, Mass.)* 3(2):143–155.

Romano, J. P., Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica* 78(1):169–211.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701. doi:10.1037/h0037350

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3):581–592. doi:10.1093/biomet/63.3.581

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine* 28(4):558–571. doi:10.1002/sim.3493

Sloczynski, T. (2013): Population average gender effects, *IZA Discussion Paper No. 7315*.

Tchetgen, E. J., Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics* 40(3):1816–1845. doi:10.1214/12-AOS990

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* 63(3):926–934. doi:10.1111/j.1541-0420.2007.00766.x

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology (Cambridge, Mass.)* 20(1):18–26. doi:10.1097/EDE.0b013e31818f69ce

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)* 21(4):540–551. doi:10.1097/EDE.0b013e3181df191c

VanderWeele, T. J., Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, Biostatistics, and Public Health* 11:1–16.

Vansteelandt, S., Bekaert, M., Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods* 1(1):129–158. doi:10.1515/2161-962X.1014

Vansteelandt, S., VanderWeele, T. J. (2012). Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics* 68(4):1019–1027. doi:10.1111/j.1541-0420.2012.01777.x