USING MISSING TYPES TO IMPROVE PARTIAL IDENTIFICATION WITH APPLICATION TO A STUDY OF HIV PREVALENCE IN MALAWI

By Zhichao Jiang and Peng Ding[1]

Princeton University and University of California, Berkeley

Frequently, empirical studies are plagued with missing data. When the data are missing not at random, the parameter of interest is not identifiable in general. Without additional assumptions, we can derive bounds of the parameters of interest, which, unfortunately, are often too wide to be informative. Therefore, it is of great importance to sharpen these worst-case bounds by exploiting additional information. Traditional missing data analysis uses only the information of the binary missing data indicator, that is, a certain data point is either missing or not. Nevertheless, real data often provide more information than a binary missing data indicator, and they often record different types of missingness. In a motivating HIV status survey, missing data may be due to the units' unwillingness to respond to the survey items or their hospitalization during the visit, and may also be due to the units' temporarily absence or relocation. It is apparent that some missing types are more likely to be missing not at random, but other missing types are more likely to be missing at random. We show that making full use of the missing types results in narrower bounds of the parameters of interest. In a real-life example, we demonstrate substantial improvement of more than 50% reduction in bound widths for estimating the prevalence of HIV in rural Malawi. As we illustrate using the HIV study, our strategy is also useful for conducting sensitivity analysis by gradually increasing or decreasing the set of types that are missing at random. In addition, we propose an easy-to-implement method to construct confidence intervals for partially identified parameters with bounds expressed as the minimums and maximums of finite parameters, which is useful for not only our problem but also many other problems involving bounds.

1. An introduction to missing data and partial identification. Missing data is a common problem for both experimental and observational studies in social and biomedical sciences. Rubin (1976) first clarified the missing at random and missing not at random mechanisms. Intuitively, the missing at random assumption requires the missing data mechanism be independent of the missing values themselves conditional on the observed data, but missing not at random allows for such dependence [Rubin (1976)]. For more subtle discussion of these concepts, see Rubin

(1976), Seaman et al. (2013), and Mealli and Rubin (2015). Missing at random is a sufficient condition to justify many missing data methods, including likelihood and Bayesian inference [Little and Rubin (2002), Yang and Kim (2016)], multiple imputation [Rubin (2004)], and inverse probability weighting and doubly robust estimation [Bang and Robins (2005), Kang and Schafer (2007)].

Unfortunately, however, the missing at random assumption is untestable and it can be too strong in practice. Without making such assumption, researchers derived bounds of the parameters of interest, considering the worst-case scenarios of the missing data [Ding and Geng (2014), Horowitz and Manski (1998), Horowitz and Manski (2000), Manski (2003), Mattei, Mealli and Pacini (2014)]. In fact, this idea had an early root in survey nonresponse, but was abandoned by its inventor Cochran (1953) because the bounds are often too wide to be useful. Recognizing the drawbacks of the extreme bounds, some researchers suggested conducting sensitivity analysis to obtain a range of estimates for the parameters corresponding to a plausible range of the sensitivity parameter [Copas and Li (1997), Molenberghs, Kenward and Goetghebeur (2001), Rotnitzky et al. (2001), Scharfstein, Rotnitzky and Robins (1999), Vansteelandt et al. (2006)]. Some researchers incorporated expert opinions [Scharfstein, Manski and Anthony (2004)] to derive narrower bounds. Other researchers imposed parametric assumptions [Miao, Ding and Geng (2016)] or used instrumental variables [Ma, Geng and Hu (2003), Shao and Wang (2016), Tang, Little and Raghunathan (2003)] to identify parameters of interest.

Motivated by a longitudinal survey of HIV prevalence in rural Malawi [Arpino, De Cao and Peracchi (2014)], we propose an alternative approach to improve the inference by exploiting the information about different missing types of the outcomes. For instance, the data have recorded that the units' HIV statuses were missing due to different reasons, that is, the outcomes have different missing types. Some of them were unwilling to respond, some of them were in a hospital, some of them were temporarily absent in the survey, some of them moved to another place to live, and some outcomes were missing due to other reasons. It is evident that some missing types may depend on the HIV status, and other missing types are very likely to be independent of the HIV status. Carefully utilizing the information of the missing types can lead to narrower bounds of the partially identified parameters compared to bounds that use only the binary missing data indicators as in the traditional analysis. Moreover, the HIV status satisfies a natural monotonicity, because a person infected at any given time point must be infected at later time points, whereas a person not infected at any given time point cannot be infected at earlier time points. Therefore, we can further improve the bounds with longitudinal HIV data. We establish theory to quantify the improvement of bounds in both cross-sectional and longitudinal data, and show that the consequential bounds of the HIV prevalence are substantially narrower in theory and in our application.

Although it is straightforward to estimate the bounds, it is challenging to construct confidence intervals for the parameters of interest. Importantly, the estimators of our bounds do not follow asymptotic normal distributions as required

by Imbens and Manski (2004) and Vansteelandt et al. (2006), and the bootstrap may lead to invalid asymptotic confidence intervals [Andrews (2000), Romano and Shaikh (2010)]. We propose a method to construct confidence intervals for partially identified parameters with upper and lower bounds expressed as the minimums and maximums of finite parameters. Our method is easy to implement, and is useful for not only our problem but also many other problems involving bound analysis in the missing data and causal inference literature [Cheng and Small (2006), Manski (2003), Mattei, Mealli and Pacini (2014), Mealli and Pacini (2013), Yang and Small (2016)].

More practically, our paper offers a novel strategy to conduct sensitivity analysis with respect to the missing data mechanism. Illustrated by the HIV study, we can gradually increase or decrease the set of types that are missing at random, and therefore obtain a sequence of results under assumptions with different restrictions on the missing data mechanism.

2. Missing data at a single time point.

2.1. *Traditional bounds with data missing not at random.* We first consider the case with a single observational time point t. Let Y_t be a binary outcome of interest. For example, in our application, Y_t is the HIV status at a given time t, with $Y_t = 1$ if infected and $Y_t = 0$ otherwise. We focus on a binary outcome, and will comment on general outcomes in Section 6. We condition on covariates implicitly in the theoretical discussion, and all conclusions hold within each stratum of covariates. Let S_t be the survival status at a given time t, with $S_t = 1$ if survive and $S_t = 0$ otherwise. The outcome Y_t is well defined only for people alive at time point t, that is, for units with $S_t = 0$, we define $Y_t = *$. In addition, Y_t is missing for some alive units. Let $\tilde{R}_t$ be the missing data indicator, with $\tilde{R}_t = 1$ if the outcome is observed, $\tilde{R}_t = 0$ if the outcome is missing, and $\tilde{R}_t = *$ if the unit is dead. In many real-world applications, an important quantity of interest is

$$\pi_t = \mathrm{pr}(Y_t = 1 \mid S_t = 1),$$

which, in the HIV example, is the prevalence of HIV of the alive people at time point t.

The pattern mixture decomposition [Little and Rubin (2002)] of the outcome distribution is

$$(2.1) \quad \begin{aligned} \pi_t = {}& \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 1, S_t = 1) \, \mathrm{pr}(\tilde{R}_t = 1 \mid S_t = 1) \\ &+ \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 0, S_t = 1) \, \mathrm{pr}(\tilde{R}_t = 0 \mid S_t = 1). \end{aligned}$$

An advantage of the above decomposition is its transparency for identification analysis. For the right-hand side of (2.1), the observed data of survivors allow for identification of $\mathrm{pr}(\tilde{R}_t = 1 \mid S_t = 1)$, $\mathrm{pr}(\tilde{R}_t = 0 \mid S_t = 1)$ and $\mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 1, S_t = 1)$, but do not contain any information about $\mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 0, S_t = 1)$

without further assumptions. Consequently, in general, we can obtain only the bounds of π_t by setting $\mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 0, S_t = 1)$ to its extreme values. Because $0 \le \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 0, S_t = 1) \le 1$, the lower and upper bounds of π_t are $\mathrm{LB}_t \le \pi_t \le \mathrm{UB}_t$, where

$$(2.2) \quad \mathrm{LB}_t = \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 1, S_t = 1)\,\mathrm{pr}(\tilde{R}_t = 1 \mid S_t = 1),$$

$$(2.3) \quad \mathrm{UB}_t = \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 1, S_t = 1)\,\mathrm{pr}(\tilde{R}_t = 1) + \mathrm{pr}(\tilde{R}_t = 0 \mid S_t = 1).$$

This type of bound analysis, considering the worst case scenarios, has a long history in statistics at least dating back to Cochran (1953)'s discussion in survey nonresponse. Similar ideas are extensively explored in econometrics [Manski (2003)]. Although the bounds in (2.2) and (2.3) do not rely on any assumptions about the missing data mechanism, they often correspond to unrealistic extreme cases that all the missing outcomes are 1 or 0, and consequently they are too wide to be useful as pointed out by Cochran (1953). Rubin (2005) echoed this view. Therefore, to make useful inference with data missing not at random, we need to exploit more information from the data and background knowledge to sharpen the bounds.

2.2. *Using nonresponse types to sharpen bounds of a single time point.* The discussion in Section 2.1 uses only the binary missing data indicator. In many applications, the data provide additional information about different types of nonresponse. For theoretical discussion, we consider a generic case with a four-valued type-specific missing data indicator. Let $R_t = 1$ if the outcome is observed, $R_t = -1$ if nonresponse is due to reasons related to the missing outcome, $R_t = 0$ if nonresponse is due to other reasons unrelated to the missing outcome, and $R_t = *$ if the unit is dead. Real problems, such as the HIV study considered in Section 5, often record many reasons of nonresponse, but we can collapse the reasons into two categories of nonresponse with $R_t = -1$ or 0. An example of the first type with $R_t = -1$ is that the unit refuses to answer the survey item about the outcome, and an example of the second type with $R_t = 0$ is that the unit moves to another place during the visit. Therefore, the coarsened binary missing data indicator $\tilde{R}_t$ equals 1 if and only if the type-specific missing data indicator R_t equals 1, $\tilde{R}_t$ equals 0 if $R_t = 0$ or -1, and $\tilde{R}_t$ equals $*$ if $R_t = *$.

The bounds in (2.2) and (2.3) do not take into account the difference in nonresponse types. We can improve them by making full use of the missing data indicator R_t and the following assumption.

ASSUMPTION 1. $\mathrm{pr}(Y_t = 1 \mid R_t = 0, S_t = 1) = \mathrm{pr}(Y_t = 1 \mid S_t = 1)$.

Assumption 1 states that among survivors, the outcome distribution of the individuals with $R_t = 0$ is the same as the outcome distribution of the whole population. According to Bayes' Theorem, Assumption 1 is equivalent to $\mathrm{pr}(R_t = 0 \mid$

$Y_t = 1, S_t = 1) = \text{pr}(R_t = 0 \mid S_t = 1)$. Therefore, Assumption 1 means that the type with $R_t = 0$ is *missing completely at random*, but it is weaker than the usual missing completely at random assumption. Recall that in the theoretical discussion, we condition on all the covariates implicitly. With covariates X_t at time t, Assumption 1 becomes $\text{pr}(Y_t = 1 \mid X_t, R_t = 0, S_t = 1) = \text{pr}(Y_t = 1 \mid X_t, S_t = 1)$, meaning that the type with $R_t = 0$ is *missing at random*. Therefore, we refer to Assumption 1 as a partial missing at random assumption.

In our motivating example, Assumption 1 means that the nonresponse corresponding to $R_t = 0$ is due to reasons unrelated to the HIV status. In the data, some individuals' HIV status is missing due to carelessness in data collection, which is purely random. Some individuals' HIV status is missing because he/she moved to another place during the survey, which is also independent of his/her HIV status. It is plausible to assume that units with these missing types constitute a simple random sample of all the units, or, equivalently, Assumption 1 holds. In addition to these reasons, there are some other reasons such as "hospitalization" and "refused to answer," which are probably related to the HIV status. As a result, we define these reasons as $R_t = -1$. Not surprisingly, Assumption 1 helps to sharpen the bounds of π_t.

THEOREM 1.   *Under Assumption 1, the sharp bounds of π_t are $\widetilde{\text{LB}}_t \leq \pi_t \leq \widetilde{\text{UB}}_t$, where*

$$\widetilde{\text{LB}}_t = \frac{\text{pr}(Y_t = 1, R_t = 1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)},$$

$$\widetilde{\text{UB}}_t = \frac{\text{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \text{pr}(R_t = -1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)}.$$

Moreover, $[\widetilde{\text{LB}}_t, \widetilde{\text{UB}}_t] \subseteq [\text{LB}_t, \text{UB}_t]$, with $[\widetilde{\text{LB}}_t, \widetilde{\text{UB}}_t] \subset [\text{LB}_t, \text{UB}_t] \subset [0, 1]$ if $\text{pr}(R_t = 0) > 0$.

Theorem 1 demonstrates that by considering different nonresponse types, we can sharpen the bounds of π_t in (2.2) and (2.3). The reduction in width of the bounds depends on $\text{pr}(R_t = 0 \mid S_t = 1)$, the proportion of units with outcomes missing at random. When $\text{pr}(R_t = 0 \mid S_t = 1) = 0$, the bounds are the same as the traditional bounds (2.2) and (2.3); when $\text{pr}(R_t = 0 \mid S_t = 1) = 1 - \text{pr}(R_t = 1 \mid S_t = 1)$, that is, $\text{pr}(R_t = -1 \mid S_t = 1) = 0$, the bounds collapse to a point, and π_t is pointly identifiable. A larger value of $\text{pr}(R_t = 0 \mid S_t = 1)$ leads to a larger reduction in the width of the bounds by taking into account the difference in nonresponse types. A practical issue of applying Theorem 1 to our real data is that we do not know whether some reasons depend on the HIV status or not. We would obtain different bounds if we define different reasons as $R_t = 0$. To deal with this issue, we will propose an approach in Section 5.4 to conducting sensitivity analysis based on Theorem 1.

Assumption 1 is weaker than missing at random, because it imposes a partial missing at random assumption that one category of nonresponse is independent of the missing outcome. Assumption 1 is similar to the identifying restrictions in pattern mixture model [Little (1993)], where the parameters for the nonrespondents are assumed to be the same as those for the respondents. In the literature, Harel and Schafer (2009) and Little, Rubin and Zangeneh (2017) discussed the role of some other partial missing at random assumptions in likelihood and Bayesian inference. We focus on nonparametric identification and bound analysis without imposing parametric modeling assumptions that are often required by likelihood and Bayesian inferences.

3. Longitudinal data missing not at random.

3.1. *Improving bounds using nonresponse types with longitudinal data.* When longitudinal data are available, the information across different time points can help to improve the bounds, as recognized by Arpino, De Cao and Peracchi (2014). For instance, HIV infection is an absorbing state, that is, a person infected at any given time point must be infected at later time points, whereas a person not infected at any given time point cannot be infected at earlier time points. Mathematically, for the people alive at both time points s and t with $s < t$, if $Y_t = 0$ then $Y_s = 0$, and if $Y_s = 1$ then $Y_t = 1$. We formally state this assumption below.

ASSUMPTION 2.   For any $s < t$, if $S_t = S_s = 1$, then $Y_s \leq Y_t$.

Assumption 2 imposes monotonicity on the individual level outcomes, which holds naturally for diseases like HIV. Under monotonicity, for $s < t$, if $Y_s = 1$, $R_s = 1$ and $S_t = 1$, then we can infer $Y_t = 1$. Note that we do not set the value of R_t to 1 in this case because the original missing mechanism contains information for the bounds. As a side note, other different forms of monotonicity are also used as identification assumptions [e.g., Angrist, Imbens and Rubin (1996), Jiang, Ding and Geng (2016), Jin and Rubin (2008)].

Under Assumption 2, we can sharpen the bounds in Theorem 1 with data at multiple time points.

THEOREM 2.   *With data at time points* $t - I, \ldots, t, \ldots, t + J$, *under Assumptions* 1 *and* 2, *the sharp bounds of* π_t *are* $\widetilde{\mathrm{LB}}_t^I \leq \pi_t \leq \widetilde{\mathrm{UB}}_t^J$ *with* $\widetilde{\mathrm{LB}}_t^I = \max\{\mathrm{L}_t(1), \mathrm{L}_t(2)\}$ *and* $\widetilde{\mathrm{UB}}_t^J = \min\{\mathrm{U}_t(1), \mathrm{U}_t(2)\}$, *where*

$$\mathrm{L}_t(1) = \widetilde{\mathrm{LB}}_t + \frac{\sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)},$$

$$\mathrm{L}_t(2) = \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1),$$

$$U_t(1) = \widetilde{UB}_t - \frac{\sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = -1, S_{t+j} = 1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)},$$

$$U_t(2) = 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1),$$

recalling that $\widetilde{LB}_t$ and $\widetilde{UB}_t$ are defined in Theorem 1.

Because people alive at time point t must be alive at time points before t, for $t' < t$, we can observe $R_{t'}$ for units with $S_t = 1$. Thus, we can calculate $\widetilde{LB}_t^I$ from the observed data. Similarly, for units with $S_{t+j} = 1$, we observe $(R_{t+j}, R_{t+j-1}, \ldots, R_t)$, and thus we can also calculate $\widetilde{UB}_t^J$ from the observed data. Note that, to obtain the bounds of π_t with longitudinal data, we require Assumption 1 holds only for time point t. Theorem 2 shows that the data at time points later than t do not have any information to improve the lower bound of π_t. Intuitively, this is because the monotonicity assumption is one-sided, that is, for a positive j, we can infer $Y_t = 0$ as long as $Y_{t+j} = 0$, but we are unsure of Y_t if $Y_{t+j} = 1$. Thus, the data at time points later than t can rule out only the possibility that some units' Y_t's take value 1 but not 0, and hence do not affect the lower bound. From a more theoretical view, with data at time points later than t, the unidentifiable term $\mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)$ can no longer attain the extreme value 1 as in Theorem 1, but it can always attain the extreme value 0, which keeps the lower bound unchanged. Similar discussion applies to the upper bound.

Comparing Theorems 1 and 2, we can see that the bounds with multiple time points are narrower than the bounds with a single time point. Therefore, collecting data at more time points can always improve the bounds, as long as Assumptions 1 and 2 hold.

In our application, monotonicity holds automatically for the HIV infection. For other diseases or other types of outcomes, this assumption may be violated. However, without monotonicity, we can still improve the bounds of Arpino, De Cao and Peracchi (2014) with Assumption 1 as shown in Theorem 1.

3.2. *Testable conditions for Assumptions* 1 *and* 2. Assumptions 1 and 2 are crucial for our theory. Therefore, it is important to check their validity empirically. Requiring the lower bound to be less than or equal to the upper bound in Theorem 2, we can obtain testable conditions implied by Assumptions 1 and 2. Moreover, because the bounds in Theorem 2 are sharp, these conditions include all testable conditions. Specifically, $\widetilde{LB}_t^I \leq \widetilde{UB}_t^J$ implies four testable inequalities. For simplicity, we give the testable conditions with two or three time points in the main text which are adequate for our empirical application. We can similarly derive the general testable conditions implied by Theorem 2, but to avoid notational complexity we give them in the Supplementary Material [Jiang and Ding (2018)].

COROLLARY 1. *With data at time points t and $t + 1$, a testable condition for Assumptions 1 and 2 is $\widetilde{\mathrm{LB}}_t \leq 1 - \mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = 0, S_t = 1)$, and it is the only testable condition.*

The condition in Corollary 1 is testable because it depends only on the distribution of the observed data, but does not depend on any missing values. Therefore, the testable condition allows us to falsify Assumptions 1 and 2 by the observed data in some scenarios. If the condition is violated, then the data invalidate the fundamental assumptions we make. We give a numerical example to illustrate Corollary 1.

EXAMPLE 1. Suppose that we have data at time points t and $t + 1$, and for simplicity, all the units are alive at both time points. Thus, we omit S_t and S_{t+1} in probabilities. Let the distributions of Y_t be $\mathrm{pr}(Y_t = 1) = 2/5$, and assume the following conditional distribution of R_t given Y_t:

$\mathbf{pr}(R_t = r \mid Y_t = y)$	$R_t = -1$	$R_t = 0$	$R_t = 1$
$Y_t = 1$	0	1/4	3/4
$Y_t = 0$	1/2	1/2	0

We further assume that the conditional distribution of (Y_{t+1}, R_{t+1}) on (Y_t, R_t) can be decomposed as $\mathrm{pr}(Y_{t+1}, R_{t+1} \mid Y_t, R_t) = \mathrm{pr}(Y_{t+1} \mid Y_t)\,\mathrm{pr}(R_{t+1} \mid R_t)$ with $\mathrm{pr}(Y_{t+1} = 1 \mid Y_t = 1) = 1$ and $\mathrm{pr}(Y_{t+1} = 1 \mid Y_t = 0) = 1/6$. Therefore, Assumption 2 holds, and $\mathrm{pr}(Y_{t+1} = 1) = 1/2$. The probabilities of the conditional distribution of R_{t+1} on R_t are

$\mathbf{pr}(R_{t+1} = r_{t+1} \mid R_t = r_t)$	$R_t = -1$	$R_t = 0$	$R_t = 1$
$R_{t+1} = 1$	1/2	1	1/2
$R_{t+1} = 0$	1/2	0	1/2
$R_{t+1} = -1$	0	0	0

Therefore, the data generating process violates Assumption 1.

From the observed data, we can verify that $\widetilde{\mathrm{LB}}_t = 1/2 > 3/8 = 1 - \mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1 \mid R_t = 0)$, which violates the condition in Corollary 1. Therefore, the observed distribution falsifies the conjunction of Assumptions 1 and 2. If we have the prior knowledge that Assumption 2 holds as in the motivating HIV example, then the observed data can falsify Assumption 1. As a result, our bounds are not applicable in this case.

Corollary 1 shows that, although the missing data mechanism in Assumption 1 cannot be tested alone, it can be tested when Assumption 2 holds a priori. For related discussion on testable conditions in other contexts, please see Balke and Pearl (1997), Cheng and Small (2006) and Kitagawa (2015).

The following corollary gives the testable condition with time points $t-1$ and t.

COROLLARY 2. *With data at time points $t-1$ and t, a testable condition for Assumptions* 1 *and* 2 *is* $\mathrm{pr}(Y_{t-1}=1, R_{t-1}=1 \mid R_t=0, S_t=1) \leq \widetilde{\mathrm{UB}}_t$, *and it is the only testable condition.*

Because our HIV example has three time points, we present theoretical results that are directly applicable. With three time points $t-1$, t and $t+1$, the sharp lower and upper bounds of π_t are $\widetilde{\mathrm{LB}}_t^1$ and $\widetilde{\mathrm{UB}}_t^1$. Thus, we have the following corollary.

COROLLARY 3. *With data at time points $t-1$, t, and $t+1$, testable conditions for Assumptions* 1 *and* 2 *are*

$$\mathrm{pr}(Y_{t+1}=0, R_{t+1}=1, S_{t+1}=1 \mid R_t=-1, S_t=1)$$
$$+ \mathrm{pr}(Y_{t-1}=1, R_{t-1}=1 \mid R_t=-1, S_t=1) \leq 1,$$
$$\mathrm{pr}(Y_{t+1}=0, R_{t+1}=1, S_{t+1}=1 \mid R_t=0, S_t=1)$$
$$+ \mathrm{pr}(Y_{t-1}=1, R_{t-1}=1 \mid R_t=0, S_t=1) \leq 1,$$
$$\mathrm{pr}(Y_{t+1}=0, R_{t+1}=1, S_{t+1}=1 \mid R_t=0, S_t=1)$$
$$+ \frac{\mathrm{pr}(Y_{t-1}=1, R_{t-1}=1, R_t=-1 \mid S_t=1)}{1-\mathrm{pr}(R_t=0 \mid S_t=1)} \leq 1 - \widetilde{\mathrm{LB}}_t,$$
$$\mathrm{pr}(Y_{t-1}=1, R_{t-1}=1 \mid R_t=0, S_t=1)$$
$$+ \frac{\mathrm{pr}(Y_{t+1}=0, R_{t+1}=1, R_t=-1, S_{t+1}=1 \mid S_t=1)}{1-\mathrm{pr}(R_t=0 \mid S_t=1)} \leq \widetilde{\mathrm{UB}}_t.$$

In practice, we should first check the testable conditions before reporting the bounds. If the conditions are violated, then the data provide evidence against Assumptions 1 and 2. In Section 5, our real data have observations at three time points and Assumption 2 holds by nature of HIV. Fortunately, the data do not contradict any testable conditions. Therefore, although the data cannot validate Assumption 1, they provide no evidence against it.

4. Confidence intervals for partially identified parameters. It is relatively easy to obtain point estimates of the lower and upper bounds by replacing the probabilities by their sample frequencies. To account for sample variability, we need to construct a confidence interval for π_t. However, because the lower and upper bounds are the minimums or maximums of some parameters, their moment estimators are not asymptotically normal. Thus, we cannot use traditional techniques without further modifications [Andrews (2000), Imbens and Manski (2004), Vansteelandt et al. (2006), Romano and Shaikh (2010)]. In this section, we propose

a method to obtain valid confidence intervals for partially identified parameters of certain forms.

4.1. *Method and theory conditioning on covariates.* We drop the subscript t for notational simplicity. In the previous sections, the parameter of interest, π, has bounds of the following form:

$$(4.1) \qquad \max\{L(1), \ldots, L(Q)\} \leq \pi \leq \min\{U(1), \ldots, U(R)\},$$

where the $L(q)$'s and $U(r)$'s are all functions of some population moments. For example, in Theorem 2, $Q = R = 2$, and $\{L(1), L(2)\}$ and $\{U(1), U(2)\}$ are the two terms in the expressions of $\widetilde{LB}_t^I$ and $\widetilde{UB}_t^J$, respectively. For any q and r, the point estimators of $L(q)$ and $U(r)$ are $\widehat{L}(q)$ and $\widehat{U}(r)$, which are asymptotically normal with means $L(q)$ and $U(r)$ and estimated standard errors $\widehat{\sigma}_L(q)$ and $\widehat{\sigma}_U(r)$.

Imbens and Manski (2004) proposed a method to construct confidence intervals for partially identified parameters, but their method requires that the estimators for the upper and lower bounds follow a joint bivariate normal distribution asymptotically. Chernozhukov, Lee and Rosen (2013) proposed a general method, but the construction of their confidence interval is nontrivial. Fortunately, the bounds in our paper have a nice feature: the lower bounds are the maximums of some parameters, and the upper bounds are the minimums of some parameters. This feature allows us to extend Imbens and Manski (2004)'s method to construct confidence intervals for π. We first review their method. They considered a simple case in which the parameter of interest has bounds $L(q) \leq \pi \leq U(r)$ for fixed q and r, and proposed to use

$$(4.2) \qquad CI(q, r) = \left[\widehat{L}(q) - C \times \widehat{\sigma}_L(q), \widehat{U}(r) + C \times \widehat{\sigma}_U(r)\right],$$

as a $(1 - \alpha)$-level confidence interval for π, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable, and C is determined by

$$(4.3) \qquad \Phi\left[C + \frac{\widehat{U}(r) - \widehat{L}(q)}{\max\{\widehat{\sigma}_L(q), \widehat{\sigma}_U(r)\}}\right] - \Phi(-C) = 1 - \alpha.$$

Note that with large samples, the solution of C in equation (4.3) is close to $\Phi^{-1}(1 - \alpha)$, the one-sided critical value based on a standard normal distribution. See Vansteelandt et al. (2006) for similar discussion.

However, the bounds in (4.1) have more complicated forms. If we know the true indices of the bounds $q_0 = \arg\max_{1 \leq q \leq Q} L(q)$ and $r_0 = \arg\min_{1 \leq r \leq R} U(r)$, then the bounds in (4.1) are simply $L(q_0) \leq \pi \leq U(r_0)$, and we can use $CI(q_0, r_0)$ as a confidence interval for π. In our case, the true values of q_0 and r_0 are unknown, we can first obtain their estimators $\widehat{q} = \arg\max_{1 \leq q \leq Q} \widehat{L}(q)$ and $\widehat{r} = \arg\min_{1 \leq r \leq R} \widehat{U}(r)$, and then use $CI(\widehat{q}, \widehat{r})$ as a confidence interval for π. In the following subsection, we will prove the validity of this new method for constructing confidence intervals.

Intuitively, as the large sample size goes to infinity, $\widehat{q}$ and $\widehat{r}$ will converge to the true values q_0 and r_0, and $\mathrm{CI}(\widehat{q}, \widehat{r})$ will then converge to $\mathrm{CI}(q_0, r_0)$. Because $\mathrm{CI}(q_0, r_0)$ has a correct asymptotic coverage rate as shown in Imbens and Manski (2004), $\mathrm{CI}(\widehat{q}, \widehat{r})$ also has a correct asymptotic coverage rate for π. The following theorem formally ensures the validity of this new confidence interval.

THEOREM 3.   *If* (1) $\{\mathrm{L}(1), \ldots, \mathrm{L}(Q)\}$ *have a unique maximum value* $\mathrm{L}(q_0)$, *and* $\{\mathrm{U}(1), \ldots, \mathrm{U}(R)\}$ *have a unique minimum value* $\mathrm{U}(r_0)$; (2) *for any q and r, the asymptotic distribution of* $\{\widehat{\mathrm{L}}(q), \widehat{\mathrm{U}}(r)\}$ *is bivariate normal with means* $\{\mathrm{L}(q), \mathrm{U}(r)\}$ *and estimated standard errors* $\{\widehat{\sigma}_{\mathrm{L}}(q), \widehat{\sigma}_{\mathrm{U}}(r)\}$, *then* $\mathrm{CI}(\widehat{q}, \widehat{r})$ *has a coverage rate at least as large as* $1 - \alpha$ *asymptotically.*

Technically, in Theorem 3 there are two types of coverage rates (the point-wise and uniform coverage rates), and they require different regularity conditions [Chernozhukov, Lee and Rosen (2013), Imbens and Manski (2004), Romano and Shaikh (2010), Vansteelandt et al. (2006)]. For simplicity we relegate the technical discussions to the online Supplementary Material. Intuitively, these conditions rule out the extreme cases that the $\mathrm{L}(q)$'s and $\mathrm{U}(r)$'s are too close. In our application, the estimated bounds are wide enough for us to believe that $\mathrm{L}(q)$'s and $\mathrm{U}(r)$'s are not close even in the presence of statistical uncertainty. Therefore, the confidence intervals calculated in our application are likely to have coverage rates at least as large as 95% asymptotically. According to Theorem 3, for each possible value of the parameter, $\mathrm{CI}(\widehat{q}, \widehat{r})$ has a coverage rate at least as large as $1 - \alpha$ asymptotically, and in the case of partial identification it is inevitable that for some values of the parameter, $\mathrm{CI}(\widehat{q}, \widehat{r})$ will have higher coverage rates than the nominal level. We give a numerical example to illustrate the procedure of Theorem 3 for constructing confidence intervals.

EXAMPLE 2.   Suppose that the parameter of interest, π, has bounds $\max\{\mathrm{L}(1), \mathrm{L}(2)\} \leq \pi \leq \min\{\mathrm{U}(1), \mathrm{U}(2)\}$. Suppose further that, from the observed data, we have point estimates $\widehat{\mathrm{L}}(1) = -0.2$, $\widehat{\mathrm{U}}(1) = 0.5$, $\widehat{\mathrm{L}}(2) = -0.1$, and $\widehat{\mathrm{U}}(2) = 0.8$, with asymptotic standard errors $\widehat{\sigma}_{\mathrm{L}}(1) = 0.01$, $\widehat{\sigma}_{\mathrm{U}}(1) = 0.04$, $\widehat{\sigma}_{\mathrm{L}}(2) = 0.02$, and $\widehat{\sigma}_{\mathrm{U}}(2) = 0.03$. Note that these numbers are artificial, not from the motivating example.

We can construct the confidence interval of π in the following steps. First, we calculate $\widehat{q} = \arg\max_{q=1,2} \mathrm{L}(q) = 2$ and $\widehat{r} = \arg\min_{r=1,2} \mathrm{U}(r) = 1$. Second, we obtain $C = 1.645$ by solving (4.3) with q and r replaced by $\widehat{q} = 2$ and $\widehat{r} = 1$. Third, we calculate $\mathrm{CI}(\widehat{q}, \widehat{r}) = \mathrm{CI}(2, 1) = [-0.133, 0.566]$ according to (4.2), which is a confidence interval for π.

The bounds with a single time point correspond to the case with $Q = R = 1$, and $\mathrm{CI}(\widehat{q}, \widehat{r})$ reduces to the one proposed by Imbens and Manski (2004). The bounds with multiple time points correspond to the case with $Q = R = 2$, and we can use

$\mathrm{CI}(\widehat{q}, \widehat{r})$ to construct a confidence interval for parameter π_t in our missing data problem.

In the Supplementary Material, we conduct simulation studies to evaluate the method for constructing confidence intervals. The simulation studies show that the coverage rates of the confidence intervals are close to the nominal levels in a wide range of situations with moderate sample sizes.

Although it is not the focus of our main paper, in the Supplementary Material, we also consider constructing confidence intervals for the bounds themselves. Because of the coverage guarantees for the bounds, the method in the Supplementary Material can be used to construct confidence intervals for the parameter of interest when condition (1) in Theorem 3 fails.

4.2. *Extension to the case adjusting for discrete covariates.* In our data, there are some discrete covariates such as gender and region, which are prognostic to the HIV status. We also find that these two covariates are related to the missing data indicator. It is therefore more reasonable to impose Assumption 1 conditioning on these covariates. In this subsection, we show how to use our method to construct confidence intervals with discrete covariates. In practice, covariates help to improve inference in three ways. First, the assumptions will generally be more plausible conditional on covariates. Second, we can first calculate the bounds conditional on the covariates and then average over them to obtain tighter bounds of the whole population [Lee (2009), Long and Hudgens (2013), Mealli and Pacini (2013)]. Third, covariates help to improve the estimation precision.

Consider a discrete covariate X corresponding to K subpopulations, with $w_k = \mathrm{pr}(X = k)$ being the proportion of subpopulation k. We fix the w_k's at the sample frequencies of X, and treat them as known constants. Therefore, our inference is conditional on the proportions of the subpopulations. The parameter of interest in subpopulation k, π_k, has bounds of the following forms:

$$(4.4) \qquad \max\{\mathrm{L}_k(1), \ldots, \mathrm{L}_k(Q_k)\} \leq \pi_k \leq \min\{\mathrm{U}_k(1), \ldots, \mathrm{U}_k(R_k)\},$$

where the subscript k is the index for the quantities of the kth subpopulation analogues of those in Section 4.1. We are interested in the overall bounds of $\pi = \sum_{k=1}^{K} w_k \pi_k$. We first consider the following simple form of bounds with known and fixed q_k's and r_k's:

$$\sum_{k=1}^{K} w_k \mathrm{L}_k(q_k) \leq \pi \leq \sum_{k=1}^{K} w_k \mathrm{U}_k(r_k).$$

If the joint distribution of the estimators of the upper and lower bounds are asymptotically normal, then we can again construct a confidence interval for π using Imbens and Manski (2004)'s method, denoted by $\mathrm{CI}(q_1, \ldots, q_k, r_1, \ldots, r_k)$. Let the true indices for the kth subpopulation bounds be $q_{k0} = \arg\max_{1 \leq q \leq Q_k} \mathrm{L}_k(q)$ and $r_{k0} = \arg\min_{1 \leq r \leq R_k} \mathrm{U}_k(r)$. They are unknown, but can be consistently

estimated by the sample analogues $\widehat{q}_{k0} = \arg\max_{1 \leq q \leq Q_k} \widehat{L}_k(q)$ and $\widehat{r}_{k0} = \arg\min_{1 \leq r \leq R_k} \widehat{U}_k(r)$. We then construct the final confidence interval for π as $\text{CI}(\widehat{q}_1, \ldots, \widehat{q}_k, \widehat{r}_1, \ldots, \widehat{r}_k)$. The following corollary extends Theorem 3, justifying the above confidence interval with a discrete covariate.

COROLLARY 4. *If* (1) *for all k*, $\{L_k(1), \ldots, L_k(Q_k)\}$ *have a unique maximum value, and* $\{U_k(1), \ldots, U_k(R_k)\}$ *have a unique minimum value;* (2) *for any q_k's and r_k's, the asymptotic distribution of* $\{\sum_{k=1}^{K} w_k \widehat{L}_k(q_k), \sum_{k=1}^{K} w_k \widehat{U}_k(r_k)\}$ *is bivariate normal, then* $\text{CI}(\widehat{q}_1, \ldots, \widehat{q}_k, \widehat{r}_1, \ldots, \widehat{r}_k)$ *has a coverage rate at least as large as* $1 - \alpha$ *asymptotically.*

For continuous covariates, we need to assume parametric or semiparametric models, and we leave this topic for future research. For time-varying covariates, we can treat them as covariates at different time points, and obtain the bounds and confidence intervals with the same procedure.

5. Application.

5.1. *Background and motivation.* Credible estimates of the prevalence of HIV are essential for policy makers to plan control programs and interventions. However, population-based surveys may be affected by missing data on respondents' HIV status [Arpino, De Cao and Peracchi (2014)]. The Malawi Diffusion and Ideational Change Project is a longitudinal survey conducted in rural Malawi every two years since 1998. The survey is from a collaboration between the University of Pennsylvania and the College of Medicine and Chancellor College at the University of Malawi. The data can be downloaded from http://www.malawi.pop.upenn.edu and include the outcomes of HIV tests for the years 2004, 2006, and 2008. We will give a data description that is relevant to our context, and refer to the original study [Anglewicz et al. (2009)] for more details. The project started with a main survey, which collected information on household structure, sexual relations, marriage, and partnership histories, etc. Starting from 2004, the Voluntary Consulting and Test survey was added to the main survey, which consisted of a short questionnaire and free tests for HIV and other sexually transmitted infections.

However, the HIV status is missing for a substantial fraction of the sample. There are different types of missing data, including temporary absence, loss of results, relocation, hospitalization, and refusal to participate. In the analysis of Arpino, De Cao and Peracchi (2014), they ignored different types of missing data, and used the binary missing data indicator to obtain the bounds of the HIV prevalence. With cross-sectional data, they calculate the worst-case bounds; with longitudinal data, they improve their bounds under monotonicity. Unfortunately, the bounds they obtained are quite wide, for example, the bound of the HIV prevalence in the south region in 2008 is [0.035, 0.529] with cross-sectional data and [0.082, 0.529] with longitudinal data. In the following analysis, we are aiming to improve the inference by taking into account the difference in missing types.

TABLE 1
Numbers of missing types by year with proportions in parentheses. The sample sizes across years differ because some units died before observations

Year	$R = -1$	$R = 0$	$R = 1$
2004	386 (9.5%)	799 (19.7%)	2877 (70.8%)
2006	323 (8.0%)	1164 (29.0%)	2531 (63.0%)
2008	453 (11.5%)	1253 (31.8%)	2233 (56.7%)

5.2. *Data description.* Following the analysis of Arpino, De Cao and Peracchi (2014), we focus on people who were interviewed in 2004, and drop units who were never successfully contacted. We use the data in years 2004, 2006, and 2008. Because HIV prevalence is defined for the population of alive individuals, we consider only the alive people when computing HIV prevalences for each of the years 2004, 2006, and 2008. The data are available from the online materials of Arpino, De Cao and Peracchi (2014). The total sample size is 4062. The HIV status is missing for 1185, 1487, and 1706 individuals in 2004, 2006, and 2008, with missing proportions 29.2%, 36.6%, and 42.0%, respectively. In the data, the reasons for the missing data in the main survey are categorized as "refused," "hospitalized," "not known," "temporarily absent," "moved." Other reasons are classified as the residual category "other," consisting mainly of people who did not fill in the questionnaire because they were too old or too sick, or for unspecified reasons that may also include migration [Arpino, De Cao and Peracchi (2014)]. The reasons for the missing data in the Voluntary Consulting and Test survey are categorized as "refused," "hospitalized," "temporarily absent," "not known," "moved," and "results lost." Other reasons for item nonresponse are classified as the category "other." The HIV status is missing if either data from the main survey or the Voluntary Consulting and Test survey are missing. In our analysis, we define $R_t = -1$ if the reason of missingness in either of the two surveys belongs to the categories "refused," "hospitalized," or "others." For the rest of the reasons of missingness, we define $R_t = 0$. Table 1 summarizes the distributions of the missing types.

In our data, there may be some reasons that are related to the HIV status but we classify them as missing at random. To deal with this, we will propose an approach to conducting sensitivity analysis in Section 5.4. We will not consider measurement error, because as stated in Arpino, De Cao and Peracchi (2014), measurement error in the two types of test (oral swabs and blood test) appears to be small.

5.3. *Data analysis.* The survey was carried out in three administrative regions: center, south, and north. Because these regions have very different demographic characteristics, we conduct analyses within subpopulations stratified by region and gender. Within each level of region and gender, it is plausible that the absence of

the individual or the loss of the individual's result does not depend on the individual's HIV status, that is, Assumption 1 holds. Because of the feature of HIV infection discussed in Section 3.1, Assumption 2 holds. Therefore, we can apply Theorems 1 and 2 to calculate the bounds of HIV prevalence. The upper and lower bounds have explicit forms, and we can estimate the bounds by replacing the probability parameters by their sample frequency analogues, and then apply the method in Section 4 to construct confidence intervals. We empirically check the testable conditions for all three years, and find that the conditions hold in all subpopulations and years. Therefore, the data provide no evidence against Assumption 1, because Assumption 2 holds automatically.
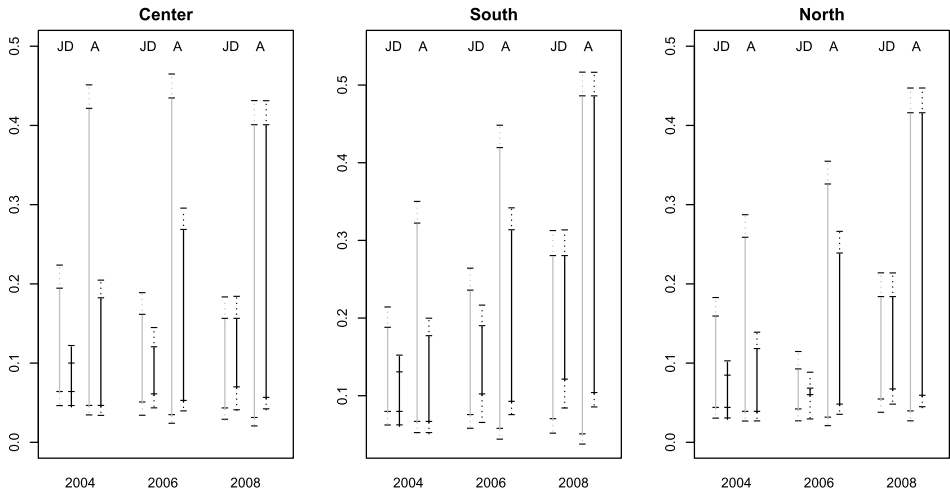
Figure 1 shows the estimated bounds and confidence intervals for the HIV prevalence from our method and Arpino, De Cao and Peracchi (2014)'s method for the three administrative regions of the country separated by gender. For descriptive convenience, we call Arpino, De Cao and Peracchi (2014)'s method the "A method" and our method the "JD method" from now on. For women in the north region in 2006 [the third plot of Figure 1(b)], the A method gives bounds $(0.047, 0.237)$, and the JD method gives bounds $(0.058, 0.067)$; the JD method reduces 95.3% of the width of the bounds. For men in the south region in 2004 [the second plot of Figure 1(a)], the A method gives bounds $(0.045, 0.172)$, and the JD method gives bounds $(0.054, 0.129)$; the JD method reduces 40.6% of the width of the bounds. The reductions in the widths of the bounds in these two subpopulations are the largest and smallest among all subpopulations, respectively. For men in the north region in 2006 [the third plot of Figure 1(a)], the A method gives a confidence interval $(0.013, 0.382)$, and the JD method gives a confidence interval $(0.019, 0.113)$; the JD method reduces 75.2% of the width of the confidence interval. For men in the south region in 2004 [the second plot of Figure 1(a)], the A method gives a confidence interval $(0.032, 0.195)$, and the JD method gives a confidence interval $(0.037, 0.154)$; the JD method reduces 28.5% of the width of the confidence interval. The reductions in the widths of the confidence intervals in these two subpopulations are the largest and smallest among all subpopulations, respectively.

Figure 2 shows the histograms of the ratios of the widths of the bounds and confidence intervals obtained by the JD method divided by those obtained by the A method. For all the results with a single time point and multiple time points, the ratios of the widths of the bounds range from 0.043 to 0.594, and the ratios of the widths of the confidence intervals range from 0.255 to 0.715. In most subpopulations, the reductions in widths of the bounds and confidence intervals are larger than 50%, demonstrating substantial improvement in inference by taking into account the different types of missing data.

We can obtain the overall bounds of men and women by first calculating the bounds in each region and then averaging the bounds over regions. Figure 3 shows
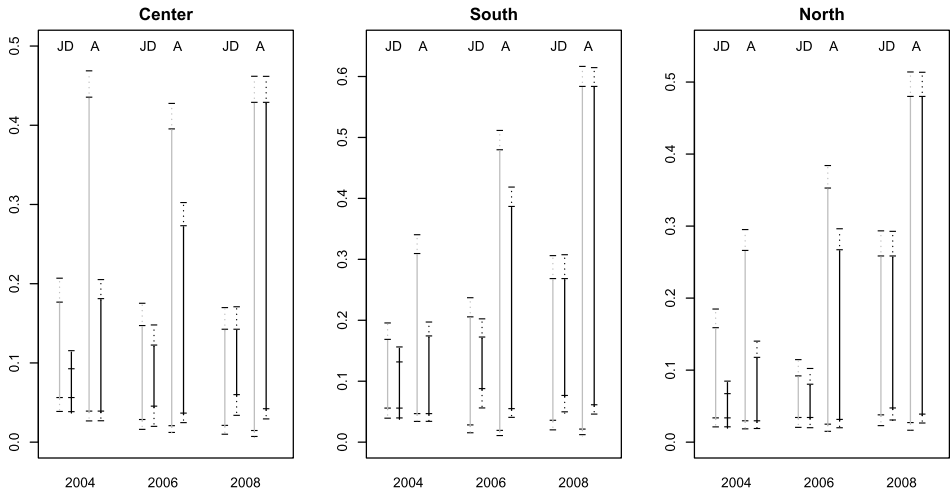
(a) Men in the three administrative regions.

(b) Women in the three administrative regions.

FIG. 1.   *The solid lines are the bounds and the dotted extended lines are the* 95% *confidence intervals for the HIV prevalence in the corresponding years. The grey lines are the results with a single time point and the black lines are the results with multiple time points. The label "JD" corresponds to the result of our method, and the label "A" corresponds to the results of Arpino, De Cao and Peracchi* (2014)*'s method.*

the overall bounds and corresponding confidence intervals using the A and JD methods. The reductions in widths of the bounds and confidence intervals are large by the JD method.
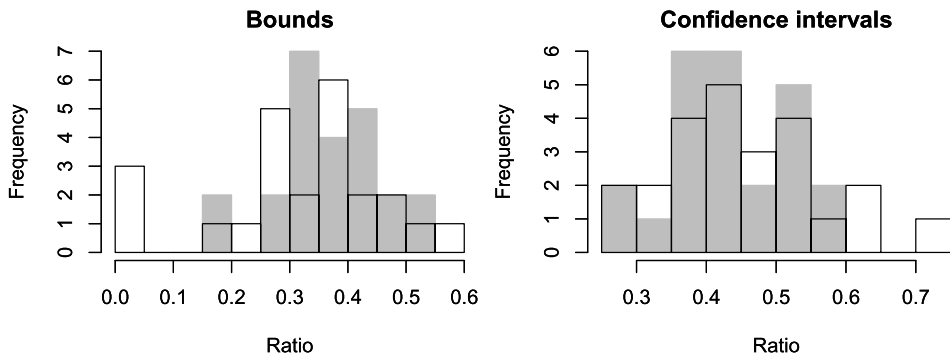
FIG. 2. *Ratios of the widths of the bounds and confidence intervals for the HIV prevalence obtained by the JD and A methods. The grey histograms are the results with a single time point and the white histograms are the results with multiple time points.*

According to Figure 1, in years 2006 and 2008, the lower bound of the HIV prevalence in Southern Malawi is the highest among the three regions for both men and women. But in 2004, the central region has the highest lower bound. This is similar for the upper bound. However, because the bounds and confidence intervals overlap, we make no attempt to draw inference about the differences in HIV prevalence across different regions. In Figure 3, the lower bounds and upper
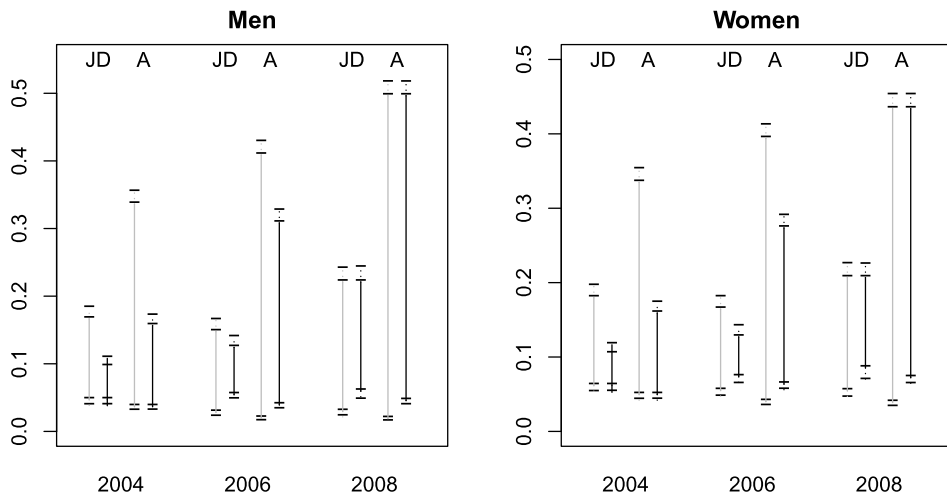

FIG. 3. *The solid lines are the bounds and the dotted extended lines are the 95% confidence intervals for the HIV prevalence in the corresponding years. The grey lines are the results with a single time point and the black lines are the results with multiple time points. The label "JD" corresponds to the result of our method, and the label "A" corresponds to the results of Arpino, De Cao and Peracchi (2014)'s method.*

bounds in 2008 are the highest among the three years for both men and women, suggesting an increasing trend of the HIV prevalence.

Note that Arpino, De Cao and Peracchi (2014) used some variables as instrumental variables or monotone instrumental variables to further narrow their bounds [Manski (2003), Manski and Pepper (2000)], which, however, invoked additional assumptions. As a future work, we can further narrow our bounds by further invoking some of these instrumental variable assumptions.

5.4. *Sensitivity analysis.*   The classification of the reasons of missingness requires prior knowledge, and sometimes it is hard to justify whether a specific reason is missing at random or not at random. For example, in our application, it is possible that absence of the individual is related to HIV status in various ways. As a result, the reason "temporarily absent" may be missing not at random, but in previous analysis, we define it as $R_t = 0$. In this case, we can conduct a sensitivity analysis by gradually increasing or decreasing the set of reasons that are missing at random. Specifically, we start with the analysis that assumes all the reasons are missing not at random. Then, we change the categories of the reasons of missingness from $R_t = -1$ to $R_t = 0$ one by one. As a result, we can see how the bounds and confidence intervals change as we strengthen the assumption about the missing types.

We conduct the sensitivity analysis for the bounds with multiple time points in 2004 in the subpopulation of women. We start with the method in Arpino, De Cao and Peracchi (2014), which classifies all the reasons of missingness as $R_t = -1$. We then change the categories of reasons "moved," "results lost or not known," and "temporarily absent" to $R_t = 0$ one by one. The results in Figure 4 show that even with a single reason classified as missing at random, we have meaningful improvement, and with three reasons classified as missing at random, we have substantial improvement. Analogously, we can conduct sensitivity analyses for other subpopulations, but we omit the results to save space.

6. Discussion.   We discussed binary outcomes in Sections 2 and 3. For general outcomes, we can dichotomize the outcomes, and apply our results for binary outcomes to obtain bounds on the distribution functions. Furthermore, bounds on the distribution function can be used to construct bounds on quantiles as suggested by Manski (2009).

In the application, there are many types of missing data. We collapsed them into two categories: one is more likely to be missing at random, and the other tends to be related to the missing outcome itself. By doing this, we do not need to model the detailed missing data patterns. Ideally, we could further distinguish the detailed missing types. But this will require more assumptions about the missing data mechanism. From the HIV data, although we know there are different missing types, we do not have very clear understanding about the missing data mechanism. If we are willing to impose parametric models, it is promising that more efficient
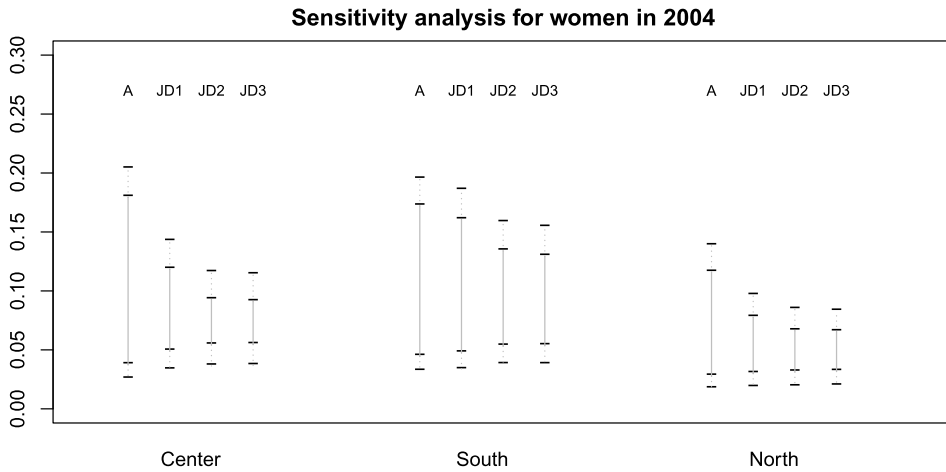
FIG. 4. *The solid lines are the bounds and the dotted extended lines are the* 95% *confidence intervals. The label "A" corresponds to the results of* Arpino, De Cao and Peracchi (2014)*'s method. The label "JD*1*" corresponds to the result of our method with* $R_t = 1$ *indicating "moved." The label "JD*2*" corresponds to the result of our method with* $R_t = 1$ *indicating "moved" and "results lost or not known." The label "JD*3*" corresponds to the result of our method with* $R_t = 1$ *indicating "moved," "results lost or not known," and "temporarily absent," which is the same as what we did in Section* 5.3.

estimates could be obtained. Our paper, however, focuses more on nonparametric and robust analyses.

The current bounds are sharp, that is, without imposing further assumptions beyond Assumptions 1 and 2, we cannot improve these bounds. In particular, we do not impose any assumptions about the transition probabilities of the missing indicator R_t. If we are willing to model the transition probability matrix, then we can obtain narrower bounds. However, from the HIV study, it does not seem obvious how to impose assumptions on this transition probability matrix. Therefore, we do not explore this direction in our current paper, but leave it to future research.

From the perspective of study designs, our results suggest that in addition to recording the binary missing indicator, researchers should also collect nonresponse types and, if possible, distinguish between types related to the outcomes and types unrelated to the outcomes.

SUPPLEMENTARY MATERIAL

Supplement to "Using missing types to improve partial identification with application to a study of HIV prevalence in Malawi" (DOI: 10.1214/17-AOAS1133SUPP; .pdf). The supplementary material consists of four parts. Section S1 gives the proofs of the theorems of the bounds. Section S2 gives the testable conditions with multiple time points. Section S3 gives the proofs of the theorem and corollary for constructing confidence interval. Section S4 shows the results of the simulation studies.

REFERENCES

ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

ANGLEWICZ, P., ADAMS, J., OBARE, F., KOHLER, H.-P. and WATKINS, S. (2009). The Malawi Diffusion and Ideational Change Project 2004–06: Data collection, data quality, and analysis of attrition. *Demogr. Res.* **20** 503–540.

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

ARPINO, B., DE CAO, E. and PERACCHI, F. (2014). Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. *J. Roy. Statist. Soc. Ser. A* **177** 587–606. MR3256525

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

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

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

CHERNOZHUKOV, V., LEE, S. and ROSEN, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737. MR3043345

COCHRAN, W. G. (1953). *Sampling Techniques*. Wiley, New York. MR0054199

COPAS, J. B. and LI, H. G. (1997). Inference for non-random samples. *J. Roy. Statist. Soc. Ser. B* **59** 55–95. MR1436555

DING, P. and GENG, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Stat. Med.* **33** 1121–1133. MR3247784

HAREL, O. and SCHAFER, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96** 37–50. MR2482133

HOROWITZ, J. L. and MANSKI, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *J. Econometrics* **84** 37–58. MR1621936

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

IMBENS, G. W. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857. MR2095534

JIANG, Z. and DING, P. (2018). Supplement to "Using missing types to improve partial identification with application to a study of HIV prevalence in Malawi." DOI:10.1214/17-AOAS1133SUPP.

JIANG, Z., DING, P. and GENG, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 829–848. MR3534352

JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. MR2463484

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

KITAGAWA, T. (2015). A test for instrument validity. *Econometrica* **83** 2043–2063. MR3414199

LEE, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.* **76** 1071–1102.

LITTLE, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

LITTLE, R. J., RUBIN, D. B. and ZANGENEH, S. Z. (2017). Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *J. Amer. Statist. Assoc.* **112** 314–320. MR3646573

LONG, D. M. and HUDGENS, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics* **69** 812–819. MR3146777

MA, W.-Q., GENG, Z. and HU, Y.-H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *J. Multivariate Anal.* **87** 24–45. MR2007260

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. Springer, New York. MR2151380

MANSKI, C. F. (2009). *Identification for Prediction and Decision*. Harvard Univ. Press, Cambridge.

MANSKI, C. F. and PEPPER, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica* **68** 997–1010. MR1771587

MATTEI, A., MEALLI, F. and PACINI, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics* **70** 278–288. MR3258033

MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131. MR3174688

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570

MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111** 1673–1683. MR3601726

MOLENBERGHS, G., KENWARD, M. G. and GOETGHEBEUR, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 15–29.

ROMANO, J. P. and SHAIKH, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica* **78** 169–211. MR2642860

ROTNITZKY, A., SCHARFSTEIN, D., SU, T.-L. and ROBINS, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57** 103–113. MR1833295

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, Hoboken, NJ. Reprint of the 1987 edition. MR2117498

RUBIN, D. B. (2005). Comment on "Multiple-bias modelling for analysis of observational data" by S. Greenland. *J. Roy. Statist. Soc. Ser. A* **168** 302.

SCHARFSTEIN, D. O., MANSKI, C. F. and ANTHONY, J. C. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the good behavior game trial. *Biometrics* **60** 154–164. MR2044111

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

SHAO, J. and WANG, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103** 175–187. MR3465829

TANG, G., LITTLE, R. J. A. and RAGHUNATHAN, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90** 747–764. MR2024755

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311

YANG, S. and KIM, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103** 244–251. MR3465836

YANG, F. and SMALL, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 299–318. MR3453657

DEPARTMENT OF POLITICS
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: zhichaoj@princeton.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: pengdingpku@berkeley.edu

Supplementary Material for "Using Missing Types to Improve Partial Identification with Application to a Study of HIV Prevalence in Malawi"

Zhichao Jiang[*] and Peng Ding[†]

The supplementary material consists of four parts.

Section S1 gives the proofs of the theorems of the bounds.

Section S2 gives the testable conditions with multiple time points.

Section S3 gives the proofs of the theorem and corollary for constructing confidence interval.

Section S4 shows the results of simulation.

S1. Proofs of the theorems of the bounds

Proof of Theorem 1. The law of total probability and Assumption 1 imply

$$
\begin{aligned}
\pi_t &= \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \mathrm{pr}(Y_t = 1 \mid R_t = 0, S_t = 1)\mathrm{pr}(R_t = 0 \mid S_t = 1) \\
&\quad + \mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)\mathrm{pr}(R_t = -1 \mid S_t = 1) \\
&= \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \pi_t \mathrm{pr}(R_t = 0 \mid S_t = 1) \\
&\quad + \mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)\mathrm{pr}(R_t = -1 \mid S_t = 1).
\end{aligned} \tag{S1}
$$

Because $0 \le \mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1) \le 1$, the decomposition (S1) implies

$$
\pi_t \ge \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \pi_t \mathrm{pr}(R_t = 0 \mid S_t = 1),
$$

$$
\pi_t \le \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \pi_t \mathrm{pr}(R_t = 0 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1).
$$

---

[*]Department of Politics, Princeton University, New Jersey 08544, USA. Email: `zhichaoj@princeton.edu`
[†]Department of Statistics, University of California, Berkeley, California, USA. Email: `pengdingpku@berkeley.edu`

Solving the above two inequalities about π_t, we have $\widetilde{\mathrm{LB}}_t \leq \pi_t \leq \widetilde{\mathrm{UB}}_t$, where

$$\widetilde{\mathrm{LB}}_t = \frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)}, \quad \widetilde{\mathrm{UB}}_t = \frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)}.$$

The lower and upper bounds can be attained when $\mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)$ equals 0 and 1, respectively. Furthermore, when $\mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)$ equals 0 or 1, the obtained distribution is compatible with Assumption 1. Thus, the bounds are sharp. It is straightforward to verify that $\widetilde{\mathrm{LB}}_t \geq \mathrm{LB}_t$, and

$$
\begin{aligned}
\mathrm{UB}_t - \widetilde{\mathrm{UB}}_t &= \mathrm{pr}(Y_t = 1 \mid \tilde{R}_t = 1, S_t = 1)\mathrm{pr}(\tilde{R}_t = 1 \mid S_t = 1) + \mathrm{pr}(\tilde{R}_t = 0 \mid S_t = 1) \\
&\quad - \frac{\mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&= \mathrm{pr}(R_t = 0 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1) - \frac{\mathrm{pr}(R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&\quad - \frac{\mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1)\mathrm{pr}(R_t = 0 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&\leq \mathrm{pr}(R_t = 0 \mid S_t = 1) - \frac{\mathrm{pr}(R_t = -1 \mid S_t = 1)\mathrm{pr}(R_t = 0 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&\quad - \frac{\mathrm{pr}(R_t = 1 \mid S_t = 1)\mathrm{pr}(R_t = 0 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&= 0,
\end{aligned}
\tag{S2}
$$

where (S2) follows from $\mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1) \leq 1$. Therefore, $[\widetilde{\mathrm{LB}}_t, \widetilde{\mathrm{UB}}_t] \subseteq [\mathrm{LB}_t, \mathrm{UB}_t]$. If $\mathrm{pr}(R_t = 0 \mid S_t = 1) > 0$, then the last inequality above holds strictly, and $[\widetilde{\mathrm{LB}}_t, \widetilde{\mathrm{UB}}_t] \subset [\mathrm{LB}_t, \mathrm{UB}_t] \subset [0, 1]$. $\qquad\square$

In the proofs below, under Assumption 1, we use the notation

$$\pi_t = \mathrm{pr}(Y_t = 1 \mid R_t = 0, S_t = 1) = \mathrm{pr}(Y_t = 1 \mid S_t = 1), \quad \theta_t = \mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1).$$

To prove Theorem 2, we need the following three lemmas.

Lemma S1. Under Assumption 1, π_t and θ_t satisfy the linear constraint

$$\pi_t \mathrm{pr}(R_t \neq 0 \mid S_t = 1) - \theta_t \mathrm{pr}(R_t = -1 \mid S_t = 1) = \mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1). \tag{S3}$$

Proof of Lemma S1. The law of total probability and Assumption 1 imply

$$
\begin{aligned}
\pi_t &= \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \pi_t \mathrm{pr}(R_t = 0 \mid S_t = 1) \\
&\quad + \mathrm{pr}(Y_t = 1 \mid R_t = -1, S_t = 1)\mathrm{pr}(R_t = -1 \mid S_t = 1) \\
&= \mathrm{pr}(Y_t = 1 \mid R_t = 1, S_t = 1)\mathrm{pr}(R_t = 1 \mid S_t = 1) + \pi_t \mathrm{pr}(R_t = 0 \mid S_t = 1) + \theta_t \mathrm{pr}(R_t = -1 \mid S_t = 1),
\end{aligned}
$$

which further implies (S3). □

Lemma S2. Under Assumption 2, we have the following constraints from the data at time points $t, \ldots, t+J$,

$$
\pi_t \leq 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1), \tag{S4}
$$

$$
\theta_t \leq 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = -1, S_t = 1). \tag{S5}
$$

Proof of Lemma S2. For $r = 0$ and -1, we can expand $\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$ with respect to the distribution of $(Y_{t+1}, R_{t+1}, S_{t+1})$:

$$
\begin{aligned}
&\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1) \\
={}& \mathrm{pr}(Y_t = 1 \mid Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid Y_{t+1} = 0, R_{t+1} = 1, S_{t+1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid R_{t+1} \neq 1, S_{t+1} = 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t+1} \neq 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid S_{t+1} = 0, R_t = r, S_t = 1)\mathrm{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1) \tag{S6} \\
={}& \mathrm{pr}(Y_t = 1 \mid Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid R_{t+1} \neq 1, S_{t+1} = 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t+1} \neq 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid S_{t+1} = 0, R_t = r, S_t = 1)\mathrm{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1) \tag{S7} \\
={}& \mathrm{pr}(Y_t = 1, Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) + \mathrm{pr}(Y_t = 1, R_{t+1} \neq 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1, S_{t+1} = 0 \mid R_t = r, S_t = 1), \tag{S8}
\end{aligned}
$$

where (S6) follows from the law of total probability, (S7) follows from $\mathrm{pr}(Y_t = 1 \mid Y_{t+1} = 0, R_{t+1} = 1, S_{t+1} = 1, R_t = r, S_t = 1) = 0$ due to monotonicity, and (S8) follows from the definition of conditional probability.

We then expand $\mathrm{pr}(Y_t = 1, R_{t+1} \neq 1, S_{t+1} = 1 \mid R_t = r, S_t = 1)$ in (S8) with respect to the

3

distribution of $(Y_{t+2}, R_{t+2}, S_{t+2})$:

$$\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$$

$$= \mathrm{pr}(Y_t = 1, Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1 \mid Y_{t+2} = 1, R_{t+2} = 1, R_{t+1} \neq 1, R_t = r, S_{t+2} = S_{t+1} = S_t = 1)$$

$$\cdot \, \mathrm{pr}(Y_{t+2} = 1, R_{t+2} = 1, R_{t+1} \neq 1, S_{t+2} = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1 \mid Y_{t+2} = 0, R_{t+2} = 1, R_{t+1} \neq 1, R_t = r, S_{t+2} = S_{t+1} = S_t = 1)$$

$$\cdot \, \mathrm{pr}(Y_{t+2} = 0, R_{t+2} = 1, R_{t+1} \neq 1, S_{t+2} = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1 \mid R_{t+2} \neq 1, R_{t+1} \neq 1, R_t = r, S_{t+2} = S_{t+1} = S_t = 1)$$

$$\cdot \, \mathrm{pr}(R_{t+2} \neq 1, R_{t+1} \neq 1, S_{t+2} = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1 \mid R_{t+1} \neq 1, R_t = r, S_{t+2} = 0, S_{t+1} = S_t = 1)$$

$$\cdot \, \mathrm{pr}(R_{t+1} \neq 1, S_{t+2} = 0, S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1, S_{t+1} = 0 \mid R_t = r, S_t = 1) \tag{S9}$$

$$= \mathrm{pr}(Y_t = 1, Y_{t+1} = 1, R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1, Y_{t+2} = 1, R_{t+2} = 1, R_{t+1} \neq 1, S_{t+2} = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1, R_{t+2} \neq 1, R_{t+1} \neq 1, S_{t+2} = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1, R_{t+1} \neq 1, S_{t+2} = 0, S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \mathrm{pr}(Y_t = 1, S_{t+1} = 0 \mid R_t = r, S_t = 1) \tag{S10}$$

where (S9) follows from the law of total probability, and (S10) follows from $\mathrm{pr}(Y_t = 1 \mid Y_{t+2} = 0, R_{t+2} = 1, R_{t+1} \neq 1, R_t = r, S_{t+2} = S_{t+1} = S_t = 1) = 0$ due to monotonicity.

Sequentially applying the law of total probability to

$$\mathrm{pr}(Y_t = 1, R_{t+j} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1), \qquad (j = 2, \ldots, J-1)$$

with respect to the distribution of $(Y_{t+j+1}, R_{t+j+1}, S_{t+j+1})$ and using the fact that

$$\mathrm{pr}(Y_t = 1 \mid Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+j} = \cdots = S_t = 1) = 0, \qquad (j = 3, \ldots, J)$$

we can further decompose (S10) as

$$
\begin{aligned}
&\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1)\\
=\ & \sum_{j=1}^{J} \mathrm{pr}(Y_t = 1 \mid Y_{t+j} = 1, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+j} = \cdots = S_t = 1)\\
&\qquad \cdot \mathrm{pr}(Y_{t+j} = 1, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)\\
&+ \mathrm{pr}(Y_t = 1 \mid R_{t+J} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+J} = \cdots = S_t = 1)\\
&\qquad \cdot \mathrm{pr}(R_{t+J} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+J} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)\\
&+ \sum_{j=2}^{J} \mathrm{pr}(Y_t = 1 \mid R_{t+j-1} \neq 1, \ldots R_{t+1} \neq 1, R_t = r, S_{t+j} = 0, S_{t+j-1} = \cdots = S_t = 1)\\
&\qquad \cdot \mathrm{pr}(R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 0, S_{t+j-1} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)\\
&+ \mathrm{pr}(Y_t = 1 \mid R_t = r, S_{t+1} = 0, S_t = 1)\mathrm{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1) \qquad\qquad \text{(S11)}
\end{aligned}
$$

where the terms in the first summation come from the expansion of $\mathrm{pr}(Y_t = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$ with respect to the event $(Y_{t+j} = 1, R_{t+j} = 1)$, and the terms in the second summation come from the expansion with respect to the event $S_{t+j} = 0$. Note that (S7) and (S10) are special cases of (S11) with $J = 1$ and $J = 2$.

Because the unobservable probabilities in (S11) are all upper bounded by 1, i.e.,

$$
\mathrm{pr}(Y_t = 1 \mid Y_{t+j} = 1, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+j} = \cdots = S_t = 1) \ \leq\ 1, \quad (j = 1, \ldots, J) \tag{S12}
$$

$$
\mathrm{pr}(Y_t = 1 \mid R_{t+J} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+J} = \cdots = S_t = 1) \ \leq\ 1, \tag{S13}
$$

$$
\mathrm{pr}(Y_t = 1 \mid R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = r, S_{t+j} = 0, S_{t+j-1} = \cdots = S_t = 1) \ \leq\ 1, \quad (j = 2, \ldots, J) \tag{S14}
$$

$$
\mathrm{pr}(Y_t = 1 \mid R_t = r, S_{t+1} = 0, S_t = 1) \ \leq\ 1. \tag{S15}
$$

we can further bound $\text{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$ by

$$\text{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$$

$$\leq \sum_{j=1}^{J} \text{pr}(Y_{t+j} = 1, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(R_{t+J} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+J} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \sum_{j=2}^{J} \text{pr}(R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 0, S_{t+j-1} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1). \tag{S16}$$

For convenience, define

$$\Delta = \text{pr}(Y_t = 1 \mid R_t = r, S_t = 1) + \sum_{j=1}^{J} \text{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = r, S_t = 1).$$

Because $S_{t'} = 1$ implies $S_t = 1$ if $t' > t$, (S16) then implies

$$\Delta \leq \sum_{j=1}^{J} \text{pr}(R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(R_{t+J} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+J} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \sum_{j=2}^{J} \text{pr}(R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 0, S_{t+j-1} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1), \tag{S17}$$

combining the two terms with summation from $j = 1$ to J. For $j = 1, \ldots, J - 1$, because $(R_{t+j} = 1, S_{t+j} = 1)$, $(R_{t+j} \neq 1, S_{t+j} = 1)$ and $S_{t+j} = 0$ consist of the whole set, we have

$$\text{pr}(R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 0, S_{t+j-1} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$+ \text{pr}(R_{t+j} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1)$$

$$= \text{pr}(R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j-1} = \cdots = S_{t+1} = 1 \mid R_t = r, S_t = 1). \tag{S18}$$

Note that the first two terms on the left-hand side of (S18) are the terms within the two summations

with index j in (S17). Sequentially applying (S18) from $j = J$ to 2 to (S17), we can obtain

$$
\begin{aligned}
\Delta \;\; &\leq \;\; \mathrm{pr}(R_{t+1} = 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) + \mathrm{pr}(R_{t+1} \neq 1, S_{t+1} = 1 \mid R_t = r, S_t = 1) \\
&\quad + \mathrm{pr}(S_{t+1} = 0 \mid R_t = r, S_t = 1) \\
&= \;\; 1.
\end{aligned}
$$

Taking $r = 0$ and -1 in $\Delta \leq 1$, we obtain

$$
\begin{aligned}
\pi_t \;\; &\leq \;\; 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1), \\
\theta_t \;\; &\leq \;\; 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = -1, S_t = 1).
\end{aligned}
$$

$\square$

Lemma S3. Under Assumption 2, we have the following constraints from the data at time points $t - I, \ldots, t$,

$$
\pi_t \;\; \geq \;\; \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1), \tag{S19}
$$

$$
\theta_t \;\; \geq \;\; \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = -1, S_t = 1). \tag{S20}
$$

Proof of Lemma S3. The proof of Lemma S3 is simpler than that of Lemma S2, because $S_t = 1$ implies that $S_{t'} = 1$ for $t' < t$.

For $r = 0$ and -1, we can expand $\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$ with respect to the distribution of

7

(Y_{t-1}, R_{t-1}):

$$\mathrm{pr}(Y_t = 1 \mid R_t = r)$$

$$= \mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 1, R_{t-1} = 1, R_t = r)\mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 0, R_{t-1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-1} = 0, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t-1} \neq 1 \mid R_t = r, S_t = 1) \tag{S21}$$

$$= \mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 0, R_{t-1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-1} = 0, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t-1} \neq 1 \mid R_t = r, S_t = 1) \tag{S22}$$

$$= \mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 0, R_{t-1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-1} = 0, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1), \tag{S23}$$

where (S21) follows from the law of total probability, (S22) follows from $\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 1, R_{t-1} = 1, R_t = r) = 1$ due to monotonicity, and (S23) follows from the definition of conditional probability.

We then expand $\mathrm{pr}(Y_t = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)$ in (S23) with respect to the distribution of (Y_{t-2}, R_{t-2}), we have

$$\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1)$$

$$= \mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 0, R_{t-1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-1} = 0, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-2} = 1, R_{t-2} = 1, R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-2} = 1, R_{t-2} = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-2} = 0, R_{t-2} = 1, R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-2} = 0, R_{t-2} = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid R_{t-2} \neq 1, R_{t-1} \neq 1, R_t = r)\mathrm{pr}(R_{t-2} \neq 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1) \tag{S24}$$

$$= \mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = r, S_t = 1) + \mathrm{pr}(Y_{t-2} = 1, R_{t-2} = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-1} = 0, R_{t-1} = 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-1} = 0, R_{t-1} = 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid Y_{t-2} = 0, R_{t-2} = 1, R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(Y_{t-2} = 0, R_{t-2} = 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)$$

$$+\mathrm{pr}(Y_t = 1 \mid R_{t-2} \neq 1, R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t-2} \neq 1, R_{t-1} \neq 1 \mid R_t = r, S_t = 1) \tag{S25}$$

where (S24) follows from the law of total probability, and (S25) follows from $\mathrm{pr}(Y_t = 1 \mid Y_{t-2} = 1, R_{t-2} = 1, R_{t-1} \neq 1, R_t = r, S_t = 1) = 1$ due to monotonicity.

Sequentially applying the law of total probability to

$$\mathrm{pr}(Y_t = 1 \mid R_{t-i} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1), \qquad (i = 2, \ldots, I-1)$$

with respect to the distribution of (Y_{t-i-1}, R_{t-i-1}) and using the fact that

$$\mathrm{pr}(Y_t = 1 \mid Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1) = 1, \qquad (i = 3, \ldots, I)$$

we can further decompose (S25) as

$$
\begin{aligned}
&\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1) \\
=\ & \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1) \\
&+ \sum_{i=1}^{I} \mathrm{pr}(Y_t = 1 \mid Y_{t-i} = 0, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1) \\
&\qquad \cdot \mathrm{pr}(Y_{t-i} = 0, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = r, S_t = 1) \\
&+ \mathrm{pr}(Y_t = 1 \mid R_{t-I} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1)\mathrm{pr}(R_{t-I} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = r, S_t = 1)
\end{aligned}
$$

$$(S26)$$

where the terms in the first summation come from the expansion of $\mathrm{pr}(Y_t = 1 \mid R_{t-i} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1)$ with respect to the event $(Y_{t-i} = 1, R_{t-i} = 1)$, and the terms in the second summation comes from the expansion with respect to the event $(Y_{t-i} = 0, R_{t-i} = 1)$. Note that (S22) and (S25) are special cases of (S11) with $J = 1$ and $J = 2$.

Because the unobservable probabilities in (S26) are all bounded below by 0, i.e.,

$$\mathrm{pr}(Y_t = 1 \mid Y_{t-i} = 0, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1) \ \geq\ 0, \quad (i = 1, \ldots, I)$$

$$(S27)$$

$$\mathrm{pr}(Y_t = 1 \mid R_{t-I} \neq 1, \ldots, R_{t-1} \neq 1, R_t = r, S_t = 1) \ \geq\ 0, \qquad (S28)$$

9

we have

$$\mathrm{pr}(Y_t = 1 \mid R_t = r, S_t = 1) \geq \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1).$$

(S29)

Taking $r = 0$ and -1, (S29) yields

$$\pi_t \;\geq\; \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1),$$

$$\theta_t \;\geq\; \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = -1, S_t = 1).$$

$\square$

Proof of Theorem 2. According to Lemmas S1–S3, we can obtain the bounds of π_t by maximizing and minimizing π_t under the constraints (S3), (S4), (S5), (S19) and (S20). These are two linear programming problems. Solving them, we obtain $\widetilde{\mathrm{LB}}_t^I \leq \pi_t \leq \widetilde{\mathrm{UB}}_t^J$, where

$$\widetilde{\mathrm{LB}}_t^I = \max \left\{ \frac{\sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} + \widetilde{\mathrm{LB}}_t, \right.$$
$$\left. \sum_{i=1}^{I} \mathrm{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1) \right\},$$

$$\widetilde{\mathrm{UB}}_t^J = \min \left\{ \widetilde{\mathrm{UB}}_t - \frac{\sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = -1, S_{t+j} = 1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)}, \right.$$
$$\left. 1 - \sum_{j=1}^{J} \mathrm{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1) \right\}.$$

The lower bound is attainable in one of the following two cases: (a) the equalities in (S27) and (S28) hold for $r = -1$ if the first term in the expression of $\widetilde{\mathrm{LB}}_t^I$ is larger than the second term; (b) the equalities in (S27) and (S28) hold for $r = 0$ if the first term in the expression of $\widetilde{\mathrm{LB}}_t^I$ is less than or equal to the second term. In either case, the distribution is compatible with Assumptions 1 and 2. Thus, the lower bound is sharp.

The upper bound is attainable in one of the following two cases: (a) the equalities in (S12)–(S15) hold for $r = -1$ if the first term in the expression of $\widetilde{\mathrm{UB}}_t^J$ is less than the second term; (b)

10

the equalities in (S12)–(S15) hold for $r = 0$ if the first term in the expression of $\widetilde{\mathrm{UB}}_t^J$ is larger than or equal to the second term. In either case, the distribution is compatible with Assumptions 1 and 2. Thus, the upper bound is sharp. □

S2. Testable conditions

Proof of Corollary 1. With time points t and $t+1$, the sharp bounds of π_t are $\widetilde{\mathrm{LB}}_t \leq \pi_t \leq \widetilde{\mathrm{UB}}_t^1$, where

$$
\begin{aligned}
&\widetilde{\mathrm{UB}}_t^1 \\
&= \min\left\{ \frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} - \frac{\mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1, R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)}, \right. \\
&\left. \qquad\quad 1 - \mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1 \mid R_t = 0, S_t = 1) \right\}
\end{aligned}
$$

Because the lower bound must be less than or equal to the upper bound, we have the constraint $\widetilde{\mathrm{LB}}_t \leq \widetilde{\mathrm{UB}}_t^1$, which is equivalent to

$$
\begin{aligned}
\frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} &\leq 1 - \mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1 \mid R_t = 0, S_t = 1), \\
\frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} &\leq \frac{\mathrm{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \mathrm{pr}(R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)} \\
&\quad - \frac{\mathrm{pr}(Y_{t+1} = 0, R_{t+1} = 1, R_t = -1 \mid S_t = 1)}{1 - \mathrm{pr}(R_t = 0 \mid S_t = 1)}.
\end{aligned}
$$

The second inequality holds automatically, and thus the first inequality is the final testable condition. □

Proof of Corollary 2. With time points $t-1$ and t, the sharp bounds of π_t are $\widetilde{\mathrm{LB}}_t^1 \leq \pi_t \leq \widetilde{\mathrm{UB}}_t$, where

$$
\widetilde{\mathrm{LB}}_t^1 = \max\left\{ \mathrm{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = 0, S_t = 1), \widetilde{\mathrm{LB}}_t + \frac{\mathrm{pr}(Y_{t-1}=1, R_{t-1}=1, R_t=-1 \mid S_t=1)}{1 - \mathrm{pr}(R_t=0 \mid S_t=1)} \right\}.
$$

Because the lower bound must be less than or equal to the upper bound, we have the constraint

11

$\widetilde{\text{LB}}_t^1 \leq \widetilde{\text{UB}}_t$, which is equivalent to

$$\text{pr}(Y_{t-1} = 1, R_{t-1} = 1 \mid R_t = 0, S_t = 1) \leq \frac{\text{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \text{pr}(R_t = -1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)},$$

$$\frac{\text{pr}(Y_t = 1, R_t = 1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)} + \frac{\text{pr}(Y_{t-1} = 1, R_{t-1} = 1, R_t = -1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)} \leq \frac{\text{pr}(Y_t = 1, R_t = 1 \mid S_t = 1) + \text{pr}(R_t = -1 \mid S_t = 1)}{1 - \text{pr}(R_t = 0 \mid S_t = 1)}.$$

The second inequality holds automatically, and thus the first inequality is the final the testable condition. $\qquad\square$

Proof of Corollary 3. Because the lower bound must be less than or equal to the upper bound, we have the constraint $\widetilde{\text{LB}}_t^1 \leq \widetilde{\text{UB}}_t^1$. The two terms in the maximum operation of $\widetilde{\text{LB}}_t^1$ must be less than or equal to the two terms in the minimum operation of $\widetilde{\text{UB}}_t^1$, which results in the four conditions of the corollary.

$\qquad\square$

With data at time points $t-I, \ldots, t, \ldots, t+J$, we can use Theorem 2 to derive testable conditions for Assumptions 1 and 2. In Theorem 2, the lower bound must be less than or equal to the upper bound. Therefore, we have the constraint $\widetilde{\text{LB}}_t^I \leq \widetilde{\text{UB}}_t^J$. The two terms in the maximum operation of $\widetilde{\text{LB}}_t^I$ must be less than or equal to the two terms in the minimum operation of $\widetilde{\text{UB}}_t^J$, which results in the following four conditions:

$$\sum_{i=1}^I \text{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = 0, S_t = 1)$$
$$\leq 1 - \sum_{j=1}^J \text{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1),$$

$$\sum_{i=1}^I \text{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1, R_t = -1 \mid S_t = 1)$$
$$\leq \text{pr}(R_t = -1 \mid S_t = 1) - \sum_{j=1}^J \text{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, R_t = -1, S_{t+j} = 1 \mid S_t = 1),$$

$$\frac{\sum_{i=1}^I \text{pr}(Y_{t-i}=1, R_{t-i}=1, R_{t-i+1}\neq 1, \ldots, R_{t-1}\neq 1, R_t=-1 \mid S_t=1)}{1 - \text{pr}(R_t=0 \mid S_t=1)} + \widetilde{\text{LB}}_t$$
$$\leq 1 - \sum_{j=1}^J \text{pr}(Y_{t+j} = 0, R_{t+j} = 1, R_{t+j-1} \neq 1, \ldots, R_{t+1} \neq 1, S_{t+j} = 1 \mid R_t = 0, S_t = 1),$$

$$\sum_{i=1}^I \text{pr}(Y_{t-i} = 1, R_{t-i} = 1, R_{t-i+1} \neq 1, \ldots, R_{t-1} \neq 1 \mid R_t = 0, S_t = 1)$$
$$\leq \widetilde{\text{UB}}_t - \frac{\sum_{j=1}^J \text{pr}(Y_{t+j}=0, R_{t+j}=1, R_{t+j-1}\neq 1, \ldots, R_{t+1}\neq 1, R_t=-1, S_{t+j}=1 \mid S_t=1)}{1 - \text{pr}(R_t=0 \mid S_t=1)}.$$

S3. Confidence intervals for partially identified parameters

We first state the regularity conditions for pointwise coverage in Theorem 3.

(R1) The width of the bounds is uniformly bounded, i.e., $C_1 < \min\{U(1), \ldots, U(R)\} - \max\{L(1), \ldots, L(Q)\} < C_2$ for some $C_1 > 0$ and $C_2 > 0$;

(R2) for any q and r, the asymptotic variance of $L(q)$ and $U(r)$, $\sigma_{\mathrm{L}}^2(q)$ and $\sigma_{\mathrm{U}}^2(r)$, are uniformly bounded, i.e., $C_3 < \sigma_{\mathrm{L}}^2(q) < C_4$ and $C_3 < \sigma_{\mathrm{U}}^2(r) < C_4$ for some $C_3 > 0$ and $C_4 > 0$.

For any q and r, let $\sigma^2(q,r)$ be the asymptotic variance of $\widehat{U}(r) - \widehat{L}(q)$. From (R2) and the Cauchy–Schwarz inequality, we have

$$\sigma^2(q,r) \le \sigma_{\mathrm{L}}^2(q) + \sigma_{\mathrm{U}}^2(r) + 2\sigma_{\mathrm{L}}(q)\sigma_{\mathrm{U}}(r) < 4C_4. \tag{S30}$$

Under the above conditions, we can show that the confidence interval $\mathrm{CI}(q_0, r_0)$ has a correct asymptotic coverage rate in the sense that $\lim_{n\to\infty} \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0)\} \ge 1 - \alpha$. We need only to verify the three conditions of Lemma 4 in Imbens and Manski (2004). In Imbens and Manski (2004)'s Lemma 4, Condition (i) follows from Condition (3) of Theorem 3, and Condition (ii) follows from (R1) and (R2). From Conditions (2) of Theorem 3, the asymptotic distribution $(\widehat{U}(r) - \widehat{L}(q)) - (U(r) - L(q))$ is normal with mean 0 and variance $\sigma^2(q, r)$. Let N denote the sample size and $\Phi(\cdot)$ denote the distribution function of the standard normal distribution. For any positive v,

$$
\begin{aligned}
&\lim_{N\to+\infty} \mathrm{pr}\left\{\sqrt{N}|\widehat{U}(r) - \widehat{L}(q) - (U(r) - L(q))| > k(U(r) - L(q))^v\right\} \\
\le\ &\lim_{N\to+\infty} \mathrm{pr}\left\{\sqrt{N}|\widehat{U}(r) - \widehat{L}(q) - (U(r) - L(q))| > kC_1^v\right\} \\
=\ &2\Phi\{-kC_1^v/\sigma(q,r)\} \\
\le\ &2\Phi\{-kC_1^v/2\sqrt{C_4}\} \xrightarrow{k\to+\infty} 0,
\end{aligned}
$$

where the first inequality follow from (R1) and the second inequality follows from (S30). Therefore, Condition (iii) in Imbens and Manski (2004)'s Lemma 4 holds.

In the following, $\mathbb{I}(\cdot)$ denotes an indicator function.

Proof of Theorem 3. We first show $\mathrm{pr}\{(\widehat{q}, \widehat{r}) \ne (q_0, r_0)\} \to 0$ as the sample size $n \to \infty$. Because of

13

the union bound $\mathrm{pr}\{(\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \leq \mathrm{pr}(\widehat{q} \neq q_0) + \mathrm{pr}(\widehat{r} \neq r_0)$, it suffices to show $\mathrm{pr}(\widehat{q} \neq q_0) \to 0$ and $\mathrm{pr}(\widehat{r} \neq r_0) \to 0$. Because of symmetry, we show only $\mathrm{pr}(\widehat{q} \neq q_0) \to 0$. Because $\widehat{L}(q) \to L(q)$ in probability for any q, the dominant convergence theorem implies

$$\mathrm{pr}(\widehat{q} \neq q_0) = \sum_{q \neq q_0} \mathrm{pr}(\widehat{q} = q) \leq \sum_{q \neq q_0} E\left[\mathbb{I}\{\widehat{L}(q) \geq \widehat{L}(q_0)\}\right] \to \sum_{q \neq q_0} E\left[\mathbb{I}\{L(q) \geq L(q_0)\}\right] = 0,$$

where the last equality holds because the uniqueness of the maximum value of $\{L(1), \ldots, L(Q)\}$ implies $L(q) < L(q_0)$ for all $q \neq q_0$.

We then show that $\mathrm{CI}(\widehat{q}, \widehat{r})$ has a correct asymptotic coverage rate. First, the confidence interval $\mathrm{CI}(q_0, r_0)$ has a correct asymptotic coverage rate in the sense that $\lim_{n \to \infty} \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0)\} \geq 1 - \alpha$ following from Imbens and Manski (2004). Second, we show that the coverage rate of $\mathrm{CI}(\widehat{q}, \widehat{r})$ is close enough to that of $\mathrm{CI}(q_0, r_0)$ asymptotically. Because

$$
\begin{aligned}
\mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0)\} &= \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0), (\widehat{q}, \widehat{r}) = (q_0, r_0)\} + \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0), (\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \\
&= \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r}), (\widehat{q}, \widehat{r}) = (q_0, r_0)\} + \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0), (\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \\
&\leq \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r})\} + \mathrm{pr}\{(\widehat{q}, \widehat{r}) \neq (q_0, r_0)\}, \quad\quad\quad\quad (\mathrm{S31})
\end{aligned}
$$

and, similarly,

$$
\begin{aligned}
\mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r})\} &= \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r}), (\widehat{q}, \widehat{r}) = (q_0, r_0)\} + \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r}), (\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \\
&= \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0), (\widehat{q}, \widehat{r}) = (q_0, r_0)\} + \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r}), (\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \\
&\leq \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0)\} + \mathrm{pr}\{(\widehat{q}, \widehat{r}) \neq (q_0, r_0)\},
\end{aligned}
$$

we have

$$|\mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r})\} - \mathrm{pr}\{\pi \in \mathrm{CI}(q_0, r_0)\}| \leq \mathrm{pr}\{(\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \to 0.$$

Therefore, this further implies $\lim_{n \to \infty} \mathrm{pr}\{\pi \in \mathrm{CI}(\widehat{q}, \widehat{r})\} \geq 1 - \alpha$. $\quad\square$

To obtain the uniform coverage property, we need more regularity conditions on the space of probability distributions $\mathcal{P}$.

(R3) For any $q_1 \neq q_2$, the distribution of $\widehat{L}(q_1) - \widehat{L}(q_2)$ converges uniformly to a normal distribution with mean $L(q_1) - L(q_2)$ and variance $\sigma_L^2(q_1, q_2)$, in the sense that for any $\epsilon > 0$, there exists

14

an M such that for all $N > M$, $P \in \mathcal{P}$ and any fixed x,

$$\left| \mathrm{pr}\left\{ \frac{\widehat{\mathrm{L}}(q_1) - \widehat{\mathrm{L}}(q_2) - (\mathrm{L}(q_1) - \mathrm{L}(q_2))}{\sigma_l(q_1, q_2)/\sqrt{N}} < x \right\} - \Phi(x) \right| < \epsilon.$$

(R4) For any $r_1 \neq r_2$, the distribution of $\widehat{\mathrm{U}}(r_1) - \widehat{\mathrm{U}}(r_2)$ converges uniformly to a normal distribution with mean $\mathrm{U}(r_1) - \mathrm{U}(r_2)$ and variance $\sigma_{\mathrm{U}}^2(r_1, r_2)$ in the same sense as (R3).

(R5) There exists a $\delta > 0$ such that for any $q_1 \neq q_2$, $|\mathrm{L}(q_1) - \mathrm{L}(q_2)| > \delta$, and for any $r_1 \neq r_2$, $|\mathrm{U}(r_1) - \mathrm{U}(r_2)| > \delta$.

Proof of Uniform Coverage. For any $q \neq q_0$ and $\epsilon > 0$, there exists an M, such that for $N > M$ and all $P \in \mathcal{P}$,

$$\begin{aligned}
&\mathrm{pr}\{\widehat{\mathrm{L}}(q_0) - \widehat{\mathrm{L}}(q) < 0\} \\
= \quad &\mathrm{pr}\left\{ \frac{\widehat{\mathrm{L}}(q_0) - \widehat{\mathrm{L}}(q) - (\mathrm{L}(q_0) - \mathrm{L}(q))}{\sigma_{\mathrm{L}}(q_0, q)/\sqrt{N}} < \frac{\mathrm{L}(q) - \mathrm{L}(q_0)}{\sigma_{\mathrm{L}}(q_1, q_2)/\sqrt{N}} \right\} \\
\leq \quad &\Phi\left\{ \frac{\mathrm{L}(q) - \mathrm{L}(q_0)}{\sigma_{\mathrm{L}}(q_1, q_2)/\sqrt{N}} \right\} + \epsilon \leq \Phi\{-\sqrt{M}\delta/2\sqrt{C_4}\} + \epsilon,
\end{aligned}$$

where the first inequality follows from Condition (R3), and the second inequality follows from $\mathrm{L}(q) - \mathrm{L}(q_0) < -\delta < 0$ ensured by Condition (R5) and $\sigma_L(q_1, q_2) \leq 2\sqrt{C_4}$ ensured by (R2). Therefore,

$$\sup_{P \in \mathcal{P}} \mathrm{pr}\{\widehat{\mathrm{L}}(q_0) - \widehat{\mathrm{L}}(q) < 0\} \leq \Phi\{-\sqrt{M}\delta/2\sqrt{C_4}\} + \epsilon.$$

We can choose M large enough to ensure $\sup_{P \in \mathcal{P}} \mathrm{pr}\{\widehat{\mathrm{L}}(q_0) - \widehat{\mathrm{L}}(q) > 0\} \leq 2\epsilon$. Because ϵ can be arbitrarily small, we have

$$\begin{aligned}
\lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \mathrm{pr}(\widehat{q} \neq q_0) &= \lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \mathrm{pr}\{\max\{\widehat{\mathrm{L}}(1), \ldots, \widehat{\mathrm{L}}(Q)\} \neq \widehat{\mathrm{L}}(q_0)\} \\
&\leq \sum_{q \neq q_0} \lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \mathrm{pr}\{\widehat{\mathrm{L}}(q_0) - \widehat{\mathrm{L}}(q) < 0\} = 0.
\end{aligned}$$

Similarly, $\lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \mathrm{pr}(\widehat{r} \neq r_0) = 0$. According to Imbens and Manski (2004)'s Lemma 4,

15

we have $\lim_{N \to +\infty} \inf_{P \in \mathcal{P}} \text{pr}\{\pi \in \text{CI}(q_0, r_0)\} \geq 1 - \alpha$. Similar to (S31), we can show that

$$
\begin{aligned}
& \lim_{N \to +\infty} \inf_{P \in \mathcal{P}} \text{pr}\{\pi \in \text{CI}(\widehat{q}, \widehat{r})\} \\
\geq \quad & \lim_{N \to +\infty} \inf_{P \in \mathcal{P}} \text{pr}\{\pi \in \text{CI}(q_0, r_0)\} - \lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \text{pr}\{(\widehat{q}, \widehat{r}) \neq (q_0, r_0)\} \\
\geq \quad & \lim_{N \to +\infty} \inf_{P \in \mathcal{P}} \text{pr}\{\pi \in \text{CI}(q_0, r_0)\} - \lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \text{pr}\{\widehat{q} \neq q_0\} - \lim_{N \to +\infty} \sup_{P \in \mathcal{P}} \text{pr}\{\widehat{r} \neq r_0\} \\
\geq \quad & 1 - \alpha.
\end{aligned}
$$

Proof of Corollary 2. The proof follows from the same argument as that of Theorem 3 above.

First, we can show that as the sample size $n \to \infty$,

$$
\text{pr}\{(\widehat{q}_1, \ldots, \widehat{q}_K, \widehat{r}_1, \ldots, \widehat{r}_K) \neq (q_{10}, \ldots, q_{K0}, r_{10}, \ldots, r_{K0})\} \to 0.
$$

Second, we have

$$
\begin{aligned}
& |\text{pr}\{\pi \in \text{CI}(\widehat{q}_1, \ldots, \widehat{q}_K, \widehat{r}_1, \ldots, \widehat{r}_K)\} - \text{pr}\{\pi \in \text{CI}(q_{10}, \ldots, q_{K0}, r_{10}, \ldots, r_{K0})\}| \\
\leq \quad & \text{pr}\{(\widehat{q}_1, \ldots, \widehat{q}_K, \widehat{r}_1, \ldots, \widehat{r}_K) \neq (q_{10}, \ldots, q_{K0}, r_{10}, \ldots, r_{K0})\} \to 0.
\end{aligned}
$$

Finally, because the "oracle" confidence interval $\text{CI}(q_{10}, \ldots, q_{K0}, r_{10}, \ldots, r_{K0})$ has correct asymptotic coverage rate, our confidence interval $\text{CI}(\widehat{q}_1, \ldots, \widehat{q}_K, \widehat{r}_1, \ldots, \widehat{r}_K)$ also has correct asymptotic coverage rate. □

The above proof is for the pointwise coverage property, and with additional regularity conditions we can establish the uniform coverage property. Due to the similarity with the discussion of Theorem 3, we omit the technical details. As a final remark, the above proofs also allow the lower bound to be the minimum of finite parameters and the upper bound to be the maximum of finite parameters. The proofs are still valid under these changes.

Extensions. In the main text, the approach for constructing confidence interval requires $\{\text{L}(1), \ldots, \text{L}(Q)\}$ have a unique maximum value $\text{L}(q_0)$ and $\{\text{U}(1), \ldots, \text{U}(R)\}$ have a unique minimum value $\text{U}(r_0)$. Here, we provide an alternative approach without this condition, which can also be used as confi-

dence intervals for the bounds themselves. First, we can rewrite the bounds as

$$L(\boldsymbol{\eta}) \le \pi \le U(\boldsymbol{\eta})$$

where $\boldsymbol{\eta}$ is identifiable from the data.

Theorem S1. For any set $C_{\boldsymbol{\eta}}$, if $\mathrm{pr}(\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}) \ge 1 - \alpha$, then $\mathrm{pr}(\pi \in C_{\pi}) \ge 1 - \alpha$, where

$$C_{\pi} = \left[\min_{\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}} L(\boldsymbol{\eta}), \boldsymbol{\eta} \max_{\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}} U(\boldsymbol{\eta}) \right].$$

Proof of Theorem S1. It follows from simple probability rules:

$$
\begin{aligned}
\mathrm{pr}(\pi \in C_{\pi}) &\ge \mathrm{pr}\left(L(\boldsymbol{\eta}) \ge \min_{\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}} L(\boldsymbol{\eta}), U(\boldsymbol{\eta}) \le \max_{\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}} U(\boldsymbol{\eta}) \right) \\
&\ge \mathrm{pr}(\boldsymbol{\eta} \in C_{\boldsymbol{\eta}}) \ge 1 - \alpha.
\end{aligned}
$$

$\square$

The forms of $U(\cdot)$, $L(\cdot)$ and $\boldsymbol{\eta}$ depend on the problems themselves, but in many cases, $\boldsymbol{\eta}$ is asymptotically Normal. Using the same logic as Theorem S1, we can derive the following results.

Corollary S1. Suppose that

$$\sqrt{N}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{d} N_k(0, \boldsymbol{\Sigma}), \quad \widehat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}.$$

Define

$$\widehat{C}_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} : N(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^{\top} \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) \le \chi^2_{k,\alpha}\}, \quad \widehat{C}_{\pi} = \left[\min_{\boldsymbol{\eta} \in \widehat{C}_{\boldsymbol{\eta}}} L(\eta), \max_{\boldsymbol{\eta} \in \widehat{C}_{\boldsymbol{\eta}}} U(\eta) \right],$$

where $\chi^2_{k,\alpha}$ is the α-quantile of the chi-squared distribution with degrees of freedom k. Then,

$$\lim_{N \to \infty} \mathrm{pr}(\pi \in \widehat{C}_{\pi}) \ge 1 - \alpha.$$

As we can see from the proof of Theorem S1, the confidence interval constructed using Corollary S1 is a confidence interval of bounds, while the approach in the main text constructs the confidence

17

intervals of the parameter of interest. As a consequence, when the conditions of Theorem 3 hold, the confidence interval constructed using Corollary S1 will be wider than that in the main text.

Similar to Section S3, additional regularity conditions on the space of probability distributions $\mathcal{P}$ will also ensure the uniform coverage property.

Corollary S2. Suppose that the distribution of $N(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})$ converges uniformly to a chi-squared distribution with degrees of freedom k, in the sense that for any $\epsilon > 0$, there exists an M such that for all $N > M$, $P \in \mathcal{P}$ and any fixed x,

$$\left| \mathrm{pr} \left\{ N(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) \le x \right\} - \chi_k^2(x) \right| < \epsilon, \tag{S32}$$

where $\chi_k^2(\cdot)$ is the cumulative distribution function of the chi-squared distribution with degrees of freedom k. Then we have

$$\lim_{N \to \infty} \inf_{P \in \mathcal{P}} \mathrm{pr}(\pi \in \widehat{C}_\pi) \ge 1 - \alpha.$$

S4. Simulation studies

S4.1. Data generating process

In this section, we conduct simulation studies to evaluate the method for constructing confidence intervals proposed in Section 4. We generate data at three time points, (Y_t, R_t) for $t = 1, 2, 3$, and calculate bounds and confidence intervals for π_t. We assume that all the units are alive at the three time points and thus we will omit S_1, S_2 and S_3 in this section. The bounds of π_2 have the form $\max\{\mathrm{L}(1), \mathrm{L}(2)\} \le \pi_2 \le \min\{\mathrm{U}(1), \mathrm{U}(2)\}$. Because the key step of our method is to replace (q, r) by $(\widehat{q}, \widehat{r})$ in $\mathrm{CI}(q, r)$, whether $(\widehat{q}, \widehat{r})$ equals the true value (q_0, r_0) will affect the coverage rates of our confidence intervals. Intuitively, the relation between $(\widehat{q}, \widehat{r})$ and (q_0, r_0) depends on the relative magnitudes between $\mathrm{L}(1)$ and $\mathrm{L}(2)$ and the relative magnitudes between $\mathrm{U}(1)$ and $\mathrm{U}(2)$. Therefore, for the lower bound, we consider three scenarios: (A1) $\mathrm{L}(1)$ is less than $\mathrm{L}(2)$; (A2) $\mathrm{L}(1)$ is larger than $\mathrm{L}(2)$; (A3) $\mathrm{L}(1)$ is very close to $\mathrm{L}(2)$. Similarly, for the upper bound, we consider three scenarios: (B1) $\mathrm{U}(1)$ is less than $\mathrm{U}(2)$; (B2) $\mathrm{U}(1)$ is larger than $\mathrm{U}(2)$; (B3) $\mathrm{U}(1)$ is very close to $\mathrm{U}(2)$. Combining the scenarios for the lower and upper bounds, we then have nine scenarios

18

covering all the possible situations.

Suppose that the complete-data likelihood factorizes as

$$\mathrm{pr}(Y_1, R_1, Y_2, R_2, Y_3, R_2) = \mathrm{pr}(Y_2, R_2)\mathrm{pr}(Y_1, R_1 \mid Y_2, R_2)\mathrm{pr}(Y_3, R_3 \mid Y_2, R_2).$$

We can first generate (Y_2, R_2), and then generate (Y_1, R_1) and (Y_3, R_3) conditional on (Y_2, R_2).

First, we generate (Y_2, R_2) from

$$\mathrm{pr}(R_2 = -1) = 1/4, \qquad \mathrm{pr}(R_2 = 0) = 1/4, \qquad \mathrm{pr}(R_2 = 1) = 1/2,$$

$$\mathrm{pr}(Y_2 = 1 \mid R_2 = -1) = 2/5, \qquad \mathrm{pr}(Y_2 = 1 \mid R_2 = 0) = 1/5, \qquad \mathrm{pr}(Y_2 = 1 \mid R_2 = 1) = 1/10.$$

Then the data generating process satisfies Assumption 1.

Second, we generate (Y_1, R_1) conditional on (Y_2, R_2). We further assume the conditional probability of (Y_1, R_1) on (Y_2, R_2) can be decomposed as $\mathrm{pr}(Y_1, R_1 \mid Y_2, R_2) = \mathrm{pr}(Y_1 \mid Y_2)\mathrm{pr}(R_1 \mid R_2)$. We set $\mathrm{pr}(Y_1 = 0 \mid Y_2 = 0) = 1$ and $\mathrm{pr}(Y_1 = 0 \mid Y_2 = 1) = 1/8$. Thus, the monotonicity $Y_1 \leq Y_2$ holds. We choose three different settings for $\mathrm{pr}(R_1 \mid R_2)$:

A1. $\mathrm{pr}(R_1 = 1 \mid R_2 = -1) = 6/7,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 0) = 0,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 1) = 5/7,$

       $\mathrm{pr}(R_1 = 0 \mid R_2 = -1) = 1/7,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 0) = 1,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 1) = 2/7;$

A2. $\mathrm{pr}(R_1 = 1 \mid R_2 = -1) = 0,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 0) = 1,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 1) = 5/7,$

       $\mathrm{pr}(R_1 = 0 \mid R_2 = -1) = 1,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 0) = 0,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 1) = 2/7;$

A3. $\mathrm{pr}(R_1 = 1 \mid R_2 = -1) = 6/7,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 0) = 5/7,$ $\mathrm{pr}(R_1 = 1 \mid R_2 = 1) = 5/7,$

       $\mathrm{pr}(R_1 = 0 \mid R_2 = -1) = 1/7,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 0) = 2/7,$ $\mathrm{pr}(R_1 = 0 \mid R_2 = 1) = 2/7.$

In A1–A3, the true data generating process implies lower bounds $\widetilde{\mathrm{LB}}_t^1 = \max(0, 1/6), \max(7/40, 1/15), \max(1/8, 1/6)$ for π_2, respectively, which correspond to the three previously mentioned scenarios of different magnitude relations between $\mathrm{L}(1)$ and $\mathrm{L}(2)$.

Third, we generate (Y_3, R_3) conditional on (Y_2, R_2). Suppose that the conditional probability of (Y_3, R_3) on (Y_2, R_2) can be decomposed as $\mathrm{pr}(Y_3, R_3 \mid Y_2, R_2) = \mathrm{pr}(Y_3 \mid Y_2)\mathrm{pr}(R_3 \mid R_2)$. We set $\mathrm{pr}(Y_3 = 1 \mid Y_2 = 1) = 1$ and $\mathrm{pr}(Y_3 = 1 \mid Y_2 = 0) = 1/8$. Thus, $Y_2 \leq Y_3$ and hence the data

generating process satisfies Assumption 2. We choose three different settings for $\mathrm{pr}(R_3 \mid R_2)$:

B1.	$\mathrm{pr}(R_3 = 1 \mid R_2 = -1) = 0,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 0) = 1,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 1) = 5/7,$
	$\mathrm{pr}(R_3 = 0 \mid R_2 = -1) = 1,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 0) = 0,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 1) = 2/7;$
B2.	$\mathrm{pr}(R_3 = 1 \mid R_2 = -1) = 4/7,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 0) = 4/7,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 1) = 5/7,$
	$\mathrm{pr}(R_3 = 0 \mid R_2 = -1) = 3/7,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 0) = 3/7,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 1) = 2/7;$
B3.	$\mathrm{pr}(R_3 = 1 \mid R_2 = -1) = 1/7,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 0) = 13/14,$	$\mathrm{pr}(R_3 = 1 \mid R_2 = 1) = 5/7,$
	$\mathrm{pr}(R_3 = 0 \mid R_2 = -1) = 6/7,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 0) = 1/14,$	$\mathrm{pr}(R_3 = 0 \mid R_2 = 1) = 2/7.$

In B1–B3, the true data generating process implies upper bounds $\widetilde{\mathrm{UB}}_t^1 = \min(3/10, 2/5), \min(3/5, 3/10),$ $\min(7/20, 3/8)$ for π_2, respectively, which correspond to the three previously mentioned scenarios of different magnitude relations between U(1) and U(2).

S4.2. Simulation results

Figure S1 shows the results of simulation studies averaged over 500 repeated samplings, with sample sizes $N = 200$ and $N = 2000$. For the nine scenarios, we obtain the estimates of the bounds, construct 95% confidence intervals for π_2, and evaluate their coverage properties over the true parameter π_2. The grey lines show that the bounds (grey solid lines) and confidence intervals (grey dotted lines) with only cross-sectional data severely over-cover the true parameters, and they are much wider than the black lines, which are the bounds (black solid lines) and intervals (black dotted lines) with longitudinal data. The black solid lines in each subfigure show that with longitudinal data, the naive bounds ignoring sample variability may under cover the true value of π_2. When $N = 200$, the coverage rates of the naive bounds are 69.6% for scenario (A2, B2) and 71.4% for scenario (A2, B3), showing the possibility of quite low coverage rates under small sample sizes. This issue becomes less severe when the sample size increases to $N = 2000$, but the coverage rates may still be below the nominal level (e.g., 93.2% for scenario (A2, B2) and 91.2% for scenario (A2, B3)). Therefore, it is necessary to account for sample variability, especially for small sample sizes.

The black dotted lines show that the confidence intervals with longitudinal data always achieve the nominal coverage rates, although the confidence intervals over cover the true value of π_2. As we mentioned in Section 4, this is because the confidence intervals for partial identified parameters

take into account not only the true values but also all other possible values of π_2. Nevertheless, the confidence intervals constructed according to our method are much narrower than the confidence intervals with only cross-sectional data.

References

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857.
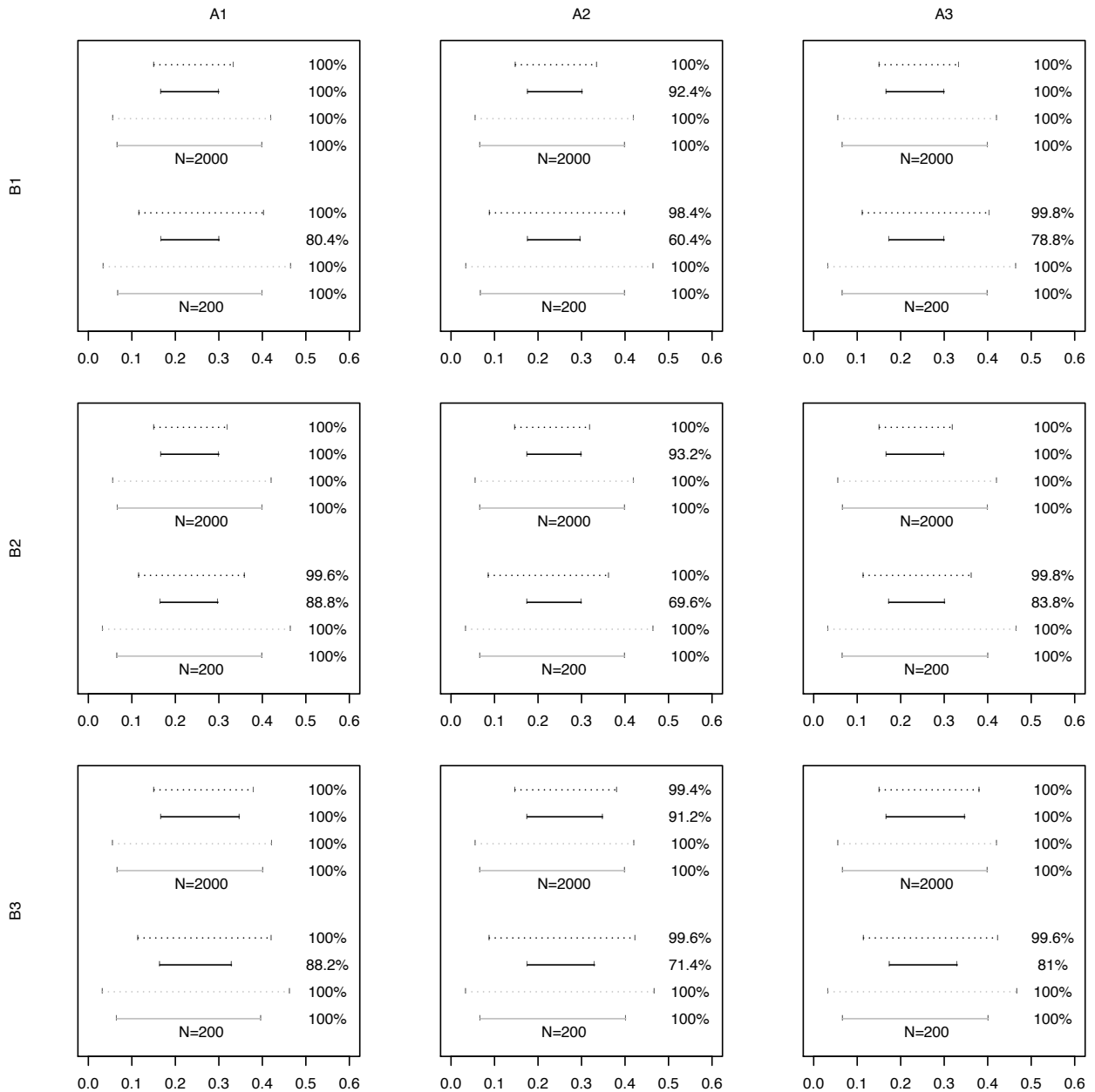
Figure S1: Simulation results of the nine combinations of parameter settings. The three columns correspond to settings A1–A3 and the three rows correspond to settings B1–B3. The grey solid and dotted lines are the bounds and confidence intervals based on the data at time point 2. The black solid and dotted lines are the bounds and confidence intervals based on the data at time points 1, 2 and 3. The bounds and confidence intervals are averaged over 500 replications, and the coverage rates are presented on the right of their corresponding lines.