Identifiability of causal effects for binary variables
with data missing due to death

Wei Yan*, Yaqin Hu, and Zhi Geng**

School of Mathematical Sciences, Peking University, Beijing, China

*email: yanwei1982@pku.edu.cn

**email: zhigeng@pku.edu.cn

SUMMARY: We discuss evaluation of causal effects with missing data due to death. Frangakis et al. (2007) proposed an approach for estimating the causal effects of interest under some assumptions. In this paper, we discuss the identifiability of the joint distribution including potential outcomes and show conditions and relaxed assumptions for identifiability. Then we propose the EM algorithms to find the maximum likelihood estimates of parameters with and without restrictions. Further we remove certain assumptions so that some parameters cannot be identified, and thus we discuss the bounds of causal effects. Our approach is evaluated via simulations, and applied to NSCOT data.

KEY WORDS: Causal inference; Missing due to death; Identifiability; EM algorithm; Bounds

1. Introduction

In many follow-up studies, we may be interested in the causal effects of a treatment on survival of individuals in different populations classified by a covariate. Values of the covariate may be missing due to death. The National Study on the Costs and Outcomes of Trauma Centers (NSCOT) in MacKenzie et al. (2006) is an example for the motivation. In the NSCOT, the treatment is the transport time (quickly or slowly) from the moment of injury in traffic accidents to arrival at the hospital, and the covariate is the activities of daily living (ADL), which evaluates a person's physical condition (poor or good). We want to evaluate the causal effects of the transport time on the survival of injured persons for the different populations classified by the ADL. Here ADL is used as a moderator to evaluate the affection to the causal effects. The transport time was recoded for every injured person, but the ADL was missing if the injured person died. Data missing due to death makes causal effects unidentifiable. Many investigators discussed the problem and proposed approaches for statistical inference with data missing due to death (Zhang and Rubin (2003), Frangakis et al. (2007), Xie and Murphy (2007)). Some assumptions on the causal mechanisms have to be made to identify the parameters of interest (Angrist, Imbens, and Rubin (1996)).

In this paper, we discuss assumptions required for identifiability of the distribution for the case with data missing due to death. Frangakis et al. (2007) proposed an approach for estimating causal effects which requires a strong monotonicity assumption that individuals are either 'protectable' by the effective treatment or else 'always survivors'. This assumption is not proper in some studies where there may be individuals of 'never survivors'. We shall relax the strong monotonicity assumption to the ordinary monotonicity and no-interaction assumptions, and we show the identifiability of the distribution with data missing due to death. We propose the EM algorithm for finding the maximum likelihood estimates (MLEs) of the parameters with restrictions under the assumptions, and further we discuss the bounds of causal effects without requiring the assumptions.

In Section 2, we describe notation and the ignorable assignment (IA) assumption which is used in the whole paper. In Section 3, we show the identifiability of the joint distribution under the strong monotonicity (SM) assumption and give the EM algorithm for finding the MLEs. In Section 4, we relax the strong monotonicity assumption to two weak assumptions and propose the EM algorithm for finding the MLEs with restrictions. In Section 5, removing these two weak assumptions and only keeping the ignorable assignment assumption, we present the significative bounds of causal effects although we cannot identify the joint distribution or causal effects. We evaluate our approach by simulations in Section 6. In Section 7, we apply our approach to the NSCOT data. Some theoretical proofs are given in Appendixes.

2. Notation

Let A be a binary covariate to denote people's health status. $A = 1$ means a poor physical condition, and $A = 0$ a better one. However, the value of A may be missing due to death. Let a binary Z denote a treatment assignment. $Z = 0$ represents a standard treatment and $Z = 1$ represents a more effective one, such as a shorter transport time. Let S^{obs} be the observed survival status: $S^{obs} = 1$ for survival, and 0 for death. An individual's S^{obs} is measured after the treatment Z is assigned. Let $S(z)$ denote the potential survival status if the individual were assigned to the treatment level z. Define $S' = (S(0), S(1))$ as a vector of potential survival outcomes. The average causal effect (ACE) of treatment Z on the survival S is defined as $E[S(1) - S(0)]$. The observed survival status $S^{obs} = S(Z)$ is one of the two potential survival outcomes $S(0)$ and $S(1)$: $S^{obs} = S(1)$ for $Z = 1$ and $S^{obs} = S(0)$ for $Z = 0$. For an individual, only one element of S' can be observed. There are four possible potential outcomes: $(1, 1)$ for 'always survivors', $(0, 1)$ for 'protectable', $(0, 0)$ for 'never survivors' and $(1, 0)$ for 'defiers'.

In causal inference, the basic *stable unit treatment value assumption* (SUTVA, Rubin (1980)) is often made, which means that there is no interference between units and only one version of

potential outcome of a certain treatment. Let $\mathbb{X}$ be a set of covariates such that the following ignorable assignment assumption holds.

ASSUMPTION 1:   *The ignorable assignment (IA) assumption.* The external factor Z is independent of (A, S') conditional on covariates $\mathbb{X}$, denoted as $Z \perp\!\!\!\perp (A, S' = (S(0), S(1))) \mid \mathbb{X}$.

For example, the IA assumption may mean that the ambulance men might decide the assignment of transport time Z only based on the injury severity or other obvious factors $\mathbb{X}$, but not the individuals' ADL A or the potential survival outcome S'. $\mathbb{X}$ can be chosen as the reasons for the remaining variability of Z.

Without loss of generality, it is assumed that the observed data are already within covariates strata $\mathbb{X} = x$, and hereafter we omit the explicit conditioning on $\mathbb{X}$ in the distributions. By the IA assumption, we can get the following factorization of the joint distribution of Z, S^{obs}, S' and A

$$P(Z, A, S', S^{obs}) = P(Z)P(A \mid Z)P(S' \mid A, Z)P(S^{obs} \mid Z, A, S')$$

$$= P(Z)P(A)P(S' \mid A)P(S^{obs} \mid Z, S'). \tag{1}$$

The second equality holds because $Z \perp\!\!\!\perp (A, S')$ and S^{obs} is completely determined by Z and S'. That is, $P(S^{obs} \mid Z, A, S') = P(S^{obs} \mid Z, S') = 1$ or 0. Although Z and S^{obs} are fully observed, A may be missing due to death and the potential outcome $S' = (S(0), S(1))$ cannot be observed completely. Thus the joint distribution is not identifiable if there are no other assumptions. In this paper, we discuss what conditions are required for the identifiability of the joint distribution (1) and present estimation approaches. When the joint distribution is identifiable, various causal effects of treatments are also identifiable, such as the average causal effect of treatment Z on survival S within level $A = a$: $E[S(1) - S(0) \mid A = a] = P(S(1) = 1 \mid A = a) - P(S(0) = 1 \mid A = a)$.

3.   Identifiability and Estimation under the Strong Monotonicity Assumption

In order to identify the joint distribution of Z, S^{obs}, S' and A, we introduce the following strong monotonicity assumption which was required in Frangakis et al. (2007).

ASSUMPTION 2:   *The strong monotonicity (SM) assumption.* Preventability of deaths from external

factor: Individuals are either 'protectable' or 'always survivors', which means $S(1) = 1$.

This SM assumption excludes 'never survivors' and 'defier' patients which is stronger than the

ordinary monotonicity assumption, i.e. $S(0) \leqslant S(1)$. This preventability, when combined with

the ignorable assignment, is testable from the observed data. Because under these assumptions,

individuals assigned to the effective treatment (i.e., $Z = 1$) must all survive, which means that all

injured persons would stay alive as long as they were transported to the hospital quickly, whereas

among those assigned to the standard treatment, some die and some survive. However, this SM

assumption may be impractical, and we relax it to two weak assumptions in Section 4.

To obtain the identifiability of the distribution in (1), we only need to show the identifiability of

$P(A)$ and $P(S'|A)$ since Z is always observed and $P(S^{obs}|Z, S')$ is determined by Z and S'. First

we show the identifiability of $P(A)$. Besides the fully observed S^{obs} and Z, by the SM assumption,

all individuals in the treatment group of $Z = 1$ are survival and thus their A is also observed.

Then we have $P(A = 1 \mid Z = 1) = P(A = 1 \mid Z = 1, S^{obs} = 1)$ which can be identified by

the proportion of individuals who have poor physical condition ($A = 1$) among those transported

to the hospital quickly ($Z = 1$). Although some values of A are missing due to death among

those individuals who are not transported to the hospital quickly, by the IA assumption, we have

$P(A = 1) = P(A = 1 \mid Z = 1)$, which can be identified by using survival individuals. Thus $P(A)$ is

identifiable.

Next we show the identifiability of $P(S' \mid A)$. By the SM assumption, we have that $P(S' = (1, S(0)) \mid A) = P(S(0) \mid A)$ and $P(S' = (0, S(0)) \mid A) = 0$. Thus we only need to show the

identifiability of $P(S(0) \mid A)$. From the independence $Z \perp\!\!\!\perp (A, S')$, we have for $a = 0$ and 1

$$P(S(0) = 0 \mid A = a) = \frac{P(S(0) = 0, A = a)}{P(A = a)}$$
$$= \frac{P(S^{obs} = 0, A = a \mid Z = 0)}{P(A = a \mid Z = 1)}$$

$$= \frac{P(A = a \mid Z = 1) - P(S^{obs} = 1, A = a \mid Z = 0)}{P(A = a \mid Z = 1)}. \tag{2}$$

It is identifiable since $P(A = a \mid Z = 1)$ can be identified under the SM assumption and $P(S^{obs} = 1, A = a \mid Z = 0) = P(A = a \mid S^{obs} = 1, Z = 0)P(S^{obs} = 1 \mid Z = 0)$ is also identifiable. Thus we show that $P(S' \mid A)$ is identifiable.

[Table 1 about here.]

[Table 2 about here.]

After showing the identifiability, we propose the EM algorithm for finding the MLEs of the distribution. The frequencies can be represented by the three-way contingency table given in Table 1 Xie and Murphy (2007). F_{zsa} denotes the frequency of individuals with values $Z = z$, $S^{obs} = s$ and $A = a$; and F_{zs+} denotes the corresponding marginal frequency of $Z = z$ and $S^{obs} = s$. '?' means an unobservable frequency. By the SM assumption, we have $F_{100} = F_{101} = 0$, that is, no individuals would die if they were carried to the hospital rapidly (i.e., $Z = 1$). In Table 2, we give the potentially completely frequencies $f_{zas_0s_1}$ which denotes the frequency with potential values $Z = z$, $A = a$, $S(0) = s_0$ and $S(1) = s_1$. Generally, the $f_{zas_0s_1}$ is unobserved since we cannot observe both $S(0)$ and $S(1)$ for any single individual. The corresponding probabilities for the potential frequencies are given in the last column of Table 2, where $p_Z = P(Z = 1)$, $p_A = P(A = 1)$ and $\theta_a = P(S' = \text{'always survivors'} \mid A = a) = P(S(0) = 1 \mid A = a)$ for $a = 0$ and 1. For Table 2, assume that the distributions of f_{0001}, f_{1011} and f_{1111} are all binomial with sizes F_{00+}, F_{110} and F_{111} respectively. Below we describe the EM algorithm. At the E-step, we find the expectations of the unobserved potential frequencies of the $(t + 1)$th iteration as follows

$$f_{0001}^{t+1} = E(f_{0001} \mid data, \theta^t) = F_{00+} \times \frac{(1 - p_A^t)(1 - \theta_0^t)}{(1 - p_A^t)(1 - \theta_0^t) + p_A^t(1 - \theta_1^t)},$$

$$f_{1011}^{t+1} = E(f_{1011} \mid data, \theta^t) = F_{110} \times \theta_0^t,$$

$$f_{1111}^{t+1} = E(f_{1111} \mid data, \theta^t) = F_{111} \times \theta_1^t,$$

where p_A^t and θ_a^t denote the estimates obtained at the tth iteration. At the M-step, the estimates of parameters can be updated as follows

$$
\begin{aligned}
p_Z^{t+1} &= \frac{F_{110} + F_{111}}{N}, \\
p_A^{t+1} &= \frac{F_{00+} - f_{0001}^{t+1} + F_{011} + F_{111}}{N}, \\
\theta_1^{t+1} &= \frac{(F_{011} + f_{1111}^{t+1})/N}{p_A^{t+1}}, \\
\theta_0^{t+1} &= \frac{(F_{010} + f_{1011}^{t+1})/N}{1 - p_A^{t+1}}.
\end{aligned}
$$

Note that the estimator of p_Z is in invariable in the iterations. The EM algorithm will be evaluated by simulation in Section 6.

4. Estimation under Relaxed Assumptions

The SM assumption of $S(1) = 1$ may be impractical in real applications. In this section, we relax it to two weaker assumptions of monotonicity and no-interaction. The three-way contingency table shown in Table 3 describes the structure of observed data without the SM assumption. Different from Table 1, the frequencies F_{100} and F_{101} are not equal to 0 but missing due to death.

[Table 3 about here.]

4.1 *Relaxed Assumptions*

We first relax the SM assumption to the monotonicity assumption and explain that it is not sufficient for identification. Then we make a no-interaction assumption to improve the identifiability. Both of them are weaker than the SM assumption.

ASSUMPTION 3: *The monotonicity (M) assumption.* Assume that $S(0) \leqslant S(1)$, which means partial preventability of deaths from external factor. That is, individuals can be stratified into three strata: 'never survivors', 'protectable' or 'always survivors'.

The monotonicity assumption allows that there may be 'never survivors' individuals who would die no matter how they are treated. This assumption may be more practical than the SM assumption. The monotonicity assumption has one more stratum $(S(0), S(1)) = (0, 0)$ than the SM assumption, and thus the relaxation induces more parameters to be estimated and causes some distributions unidentifiable. The three strata 'never survivors', 'protectable' and 'always survivors' are denoted by their initial letters: 'n', 'p' and 'a'. The probabilities of these principal strata can be identified as follows

$$P(S'=\text{n})=P(S(1) = 0) = P(S^{obs} = 0 \mid Z = 1),$$

$$P(S'=\text{a})=P(S(0) = 1) = P(S^{obs} = 1 \mid Z = 0),$$

$$P(S'=\text{p})=1-P(S'=\text{n})-P(S'=\text{a})=P(S^{obs}= 1 \mid Z=1)-P(S^{obs}= 1 \mid Z=0).$$

From the above formulas, we can see that the identifiability of $P(S' \mid A)$ and $P(S(z) \mid A)$ is equivalent under the monotonicity assumption. In addition, by $Z \perp\!\!\!\perp (A, S')$, the distribution of A in the 'protectable' stratum can be identified by

$$
\begin{aligned}
P(A = 1 \mid S' = \text{p}) &= \frac{P(A = 1)-P(A = 1, S' = \text{n})-P(A = 1, S' = \text{a})}{P(S' = \text{p})} \\
&= \frac{P(A = 1 \mid Z = 1)-P(A = 1, S'= \text{n} \mid Z = 1)-P(A = 1, S'= \text{a} \mid Z = 0)}{P(S^{obs} = 1 \mid Z = 1) - P(S^{obs} = 1 \mid Z = 0)} \\
&= \frac{P(A = 1, S^{obs} = 1 \mid Z = 1) - P(A = 1, S^{obs} = 1 \mid Z = 0)}{P(S^{obs} = 1 \mid Z = 1) - P(S^{obs} = 1 \mid Z = 0)}.
\end{aligned}
$$

For the 'always survivors' stratum, we can identify the distribution of A as follows

$$P(A = 1 \mid S' = \text{a}) = P(A = 1 \mid S(0) = 1) = P(A = 1 \mid S(0) = 1, Z = 0)$$

$$= P(A = 1 \mid S^{obs} = 1, Z = 0).$$

But the distribution of A in the 'never survivors' stratum cannot be identified because A is always missing for this stratum. Furthermore, this means that we cannot identify $P(A)$ or $P(S' \mid A)$. In order to improve identifiability, we propose the following no-interaction assumption.

ASSUMPTION 4: *The no-interaction (NI) assumption*. The odds ratios (OR) of survival status with respect to physical condition in both standard and effective treatments are the same, that is $OR_{S^{obs}A \mid Z=1} =$

$OR_{S^{obs}A|Z=0}$, which is equivalent to

$$\frac{P(S^{obs} = 1, A = 1 \mid Z = 1)P(S^{obs} = 0, A = 0 \mid Z = 1)}{P(S^{obs} = 1, A = 0 \mid Z = 1)P(S^{obs} = 0, A = 1 \mid Z = 1)}$$
$$= \frac{P(S^{obs} = 1, A = 1 \mid Z = 0)P(S^{obs} = 0, A = 0 \mid Z = 0)}{P(S^{obs} = 1, A = 0 \mid Z = 0)P(S^{obs} = 0, A = 1 \mid Z = 0)}.$$

The no-interaction assumption means that the physical condition A has the same association with the survival S conditional on the treatment Z. The assumption holds for the logistic regression of S^{obs} on A and Z which does not include the interaction between A and Z. Also the no-interaction assumption is implied by the SM assumption since the SM assumption means $F_{100} = F_{101} = 0$ which implies the no-interaction assumption of $F_{111}F_{100}F_{010}F_{001} = F_{110}F_{101}F_{011}F_{000}$.

4.2 *Identifiability*

We first prove the identifiability of the full joint distribution (1) under the assumptions which are weaker than the SM assumption.

THEOREM 1: *The joint distribution of Z, S^{obs}, S' and A can be identified under the IA, M and NI assumptions.*

This theorem means that the joint distribution of the full variables can be determined uniquely by the distribution of incompletely observed variables. For the contingency table in Table 3, we have four unobserved frequencies $\{F_{z0a}$, for z, $a = 0$, $1\}$, and others are observable. Given the observed marginal frequencies F_{00+} and F_{10+}, two of the unobserved frequencies are free. From the IA and NI assumptions, we have two equations, and thus we can solve these equations to get $\{F_{z0a}\}$. Further under the monotonicity assumption, we can identify the joint distribution of Z, S^{obs}, S' and A. For example, from (1), we have $P(Z = 1, A = 0, S' = \text{n}, S^{obs} = 0) = P(Z = 1)P(A = 0)P(S(1) = 0 \mid A = 0) = P(Z = 1)P(A = 0)P(S^{obs} = 0 \mid A = 0, Z = 1)$, which can be identified by $(F_{1++}/N) \times (F_{++0}/N) \times (F_{100}/F_{1+0})$.

[Table 4 about here.]

The structure of potentially complete data with principal strata of survival S' is illustrated in the first five columns of Table 4. The potential frequency $f_{zas_0s_1}$ denotes the frequencies completely categorized by the all potential variables, and the corresponding probabilities are given in the last column. The SM assumption restricts that the stratum $S' = (0,0)$ is empty, but the monotonicity assumption does not. For simplifying the notation, we define the parameters $p_Z = P(Z = 1)$, $p_A = P(A = 1)$, $\theta_{s'a} = P(S' = s' \mid A = a)$ for $s' = $ n, p, a and for $a = 0, 1$, where $\theta_{n1} + \theta_{p1} + \theta_{a1} = \theta_{n0} + \theta_{p0} + \theta_{a0} = 1$. The joint distribution (1) can be expressed by these parameters. The proof of Theorem 1 in Appendix does not only show the identifiability, but also gives a method for the moment estimation of these parameters.

4.3 *The EM algorithm*

In this subsection, we give the EM algorithm for finding the MLEs of all parameters in Table 4. In the M-step of our approach, the problem of finding MLEs of multiple parameters with restrictions is reduced to a problem of maximizing an objective function with a single parameter and without restrictions. Assuming that all probabilities are positive, from the no-interaction assumption, we can have

$$\frac{P(S^{obs} = 1, A = 1, Z = 1)P(S^{obs} = 0, A = 0, Z = 1)}{P(S^{obs} = 1, A = 0, Z = 1)P(S^{obs} = 0, A = 1, Z = 1)}$$
$$= \frac{P(S^{obs} = 1, A = 1, Z = 0)P(S^{obs} = 0, A = 0, Z = 0)}{P(S^{obs} = 1, A = 0, Z = 0)P(S^{obs} = 0, A = 1, Z = 0)}$$
$$\Leftrightarrow \frac{p_Z p_A(1-\theta_{n1}) \cdot p_Z(1-p_A)\theta_{n0}}{p_Z(1-p_A)(1-\theta_{n0}) \cdot p_Z p_A \theta_{n1}} = \frac{(1-p_Z)(1-p_A)(1-\theta_{a0}) \cdot (1-p_Z)p_A \theta_{a1}}{(1-p_Z)(1-p_A)\theta_{a0} \cdot (1-p_Z)p_A(1-\theta_{a1})}$$
$$\Leftrightarrow \theta_{n0}\theta_{a0}(1-\theta_{n1})(1-\theta_{a1}) = \theta_{n1}\theta_{a1}(1-\theta_{n0})(1-\theta_{a0}),$$

and finally we get a restriction $\theta_{n1}\theta_{a1}\theta_{p0} = \theta_{p1}\theta_{n0}\theta_{a0}$. For the potentially complete data in Table 4, from the joint distribution (1), we have the likelihood function

$$L(p_Z, p_A, \theta_{n1}, \theta_{p1}, \theta_{a1}, \theta_{n0}, \theta_{p0}, \theta_{a0})$$

$$\propto p_Z^{N_{Z_1}}(1-p_Z)^{N_{Z_0}}p_A^{N_{A_1}}(1-p_A)^{N_{A_0}}\theta_{n1}^{N_{n1}}\theta_{p1}^{N_{p1}}\theta_{a1}^{N_{a1}}\theta_{n0}^{N_{n0}}\theta_{p0}^{N_{p0}}\theta_{a0}^{N_{a0}},$$

where $N_{Z_z} = \#\{Z = z\}$ denotes the frequency of $Z = z$, $N_{s'a} = \#\{S' = s', A = a\} = f_{0as_0s_1} + f_{1as_0s_1}$

denotes the frequency of $S' = s' = (s_0, s_1)$ and $A = a$ in the whole population, and $N_{A_a} = N_{na} +$

$N_{pa} + N_{aa}$. Since the parameters (p_Z, p_A) are distinct to the parameters $\theta = (\theta_{n1}, \theta_{p1}, \theta_{a1}, \theta_{n0}, \theta_{p0}, \theta_{a0})$,

the likelihood function can be factorized into $L(p_Z, p_A, \theta) \propto L(p_Z, p_A)L(\theta)$. Thus we can find MLEs

of these parameters separately, and we have the MLE $\hat{p}_Z = N_{Z_1}/N$ and $\hat{p}_A = N_{A_1}/N$. The likelihood

function of the parameters θ is

$$L(\theta) \propto \theta_{n1}^{N_{n1}}\theta_{p1}^{N_{p1}}\theta_{a1}^{N_{a1}}\theta_{n0}^{N_{n0}}\theta_{p0}^{N_{p0}}\theta_{a0}^{N_{a0}}.$$

The parameters θ should satisfy the restrictions $\theta_{n1} + \theta_{p1} + \theta_{a1} = \theta_{n0} + \theta_{p0} + \theta_{a0} = 1$ and $\theta_{n1}\theta_{a1}\theta_{p0} =$

$\theta_{p1}\theta_{n0}\theta_{a0}$. Using the Lagrange multiplier method, the objective function is $l = \log L(\theta) + \lambda(\theta_{n1}\theta_{a1}\theta_{p0} -$

$\theta_{p1}\theta_{n0}\theta_{a0})$ and the restriction $\theta_{n1} + \theta_{p1} + \theta_{a1} = \theta_{n0} + \theta_{p0} + \theta_{a0} = 1$ is considered in the calculation.

The EM algorithm for finding the MLEs under the restrictions can be described as follows.

E-step: Compute the expectations conditional on the observed data and the current estimates of

parameters. In the $(t + 1)$th iteration, the incomplete data F_{000}, F_{100}, f_{0000}, f_{0100}, f_{1001} and f_{1101} in

Table 4 can be estimated by

$$E(F_{000} \mid data, \theta^t) = F_{00+} \times \frac{(1-p_A^t)(\theta_{n0}^t + \theta_{p0}^t)}{(1-p_A^t)(\theta_{n0}^t + \theta_{p0}^t) + p_A^t(\theta_{n1}^t + \theta_{p1}^t)},$$

$$E(F_{100} \mid data, \theta^t) = F_{10+} \times \frac{(1-p_A^t)\theta_{n0}^t}{(1-p_A^t)\theta_{n0}^t + p_A^t\theta_{n1}^t},$$

$$E(f_{0000} \mid data, \theta^t) = E(F_{000} \mid data, \theta^t) \times \frac{\theta_{n0}^t}{\theta_{n0}^t + \theta_{p0}^t},$$

$$E(f_{0100} \mid data, \theta^t) = (F_{00+} - E(F_{000} \mid data, \theta^t)) \times \frac{\theta_{n1}^t}{\theta_{n1}^t + \theta_{p1}^t},$$

$$E(f_{1001} \mid data, \theta^t) = F_{110} \times \frac{\theta_{p0}^t}{1 - \theta_{n0}^t},$$

$$E(f_{1101} \mid data, \theta^t) = F_{111} \times \frac{\theta_{p1}^t}{1 - \theta_{n1}^t},$$

where θ^t denotes the current estimate from the M-step in the tth iteration.

M-step: Maximize the likelihood function $L(\theta)$. The frequencies $\{N_{s'a}\}$ for the likelihood function

$L(\theta)$ can be obtained according to Table 4. For simplicity, we omit the superscript t of N^t which denotes the tth iteration. We consider the following two cases separately:

(1) For the case of $N_{n1}N_{a1}N_{p0} = N_{p1}N_{n0}N_{a0} = 0$, it is obvious that $\theta_{s'a}^{t+1} = N_{s'a}/N_{A_a}$ for $s' = $ n, p, a and $a = 0, 1$ maximize $L(\theta)$, and they satisfy the restriction $\theta_{n1}\theta_{a1}\theta_{p0} = \theta_{p1}\theta_{n0}\theta_{a0}$.

(2) Otherwise for the case of $N_{n1}N_{a1}N_{p0} \neq 0$ or $N_{p1}N_{n0}N_{a0} \neq 0$, we find the partial derivatives of l with respect to each $\theta_{s'a}$ and set them to 0. Then we obtain the following estimation equations

$$\lambda\theta_{n1}\theta_{a1}\theta_{p0} = \lambda\theta_{p1}\theta_{n0}\theta_{a0} = \frac{N_{p1}\theta_{n1} - N_{n1}\theta_{p1}}{\theta_{n1} + \theta_{p1}} = \frac{N_{n0}\theta_{p0} - N_{p0}\theta_{n0}}{\theta_{n0} + \theta_{p0}},$$

$$\frac{N_{n1} + \lambda\theta_{n1}\theta_{a1}\theta_{p0}}{\theta_{n1}} = \frac{N_{p1} - \lambda\theta_{n1}\theta_{a1}\theta_{p0}}{\theta_{p1}} = \frac{N_{a1} + \lambda\theta_{n1}\theta_{a1}\theta_{p0}}{\theta_{a1}},$$

$$\frac{N_{n0} + \lambda\theta_{p1}\theta_{n0}\theta_{a0}}{\theta_{n0}} = \frac{N_{p0} - \lambda\theta_{p1}\theta_{n0}\theta_{a0}}{\theta_{p0}} = \frac{N_{a0} + \lambda\theta_{p1}\theta_{n0}\theta_{a0}}{\theta_{a0}}.$$

From the above formulas, we can express $\theta_{p1}, \theta_{a1}, \theta_{n0}, \theta_{p0}, \theta_{a0}$ by the following functions of θ_{n1}

$$\theta_{p1} = -\frac{N_{A_1} + N_{p1}}{N_{p1} + N_{a1}}\left(\theta_{n1} - \frac{N_{n1} + N_{p1}}{N_{A_1} + N_{p1}}\right),$$

$$\theta_{a1} = \frac{N_{n1} + N_{p1}}{N_{p1} + N_{a1}}\left(\theta_{n1} - \frac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}}\right),$$

$$\theta_{n0} = \frac{N_{A_1} + N_{n0}}{N_{A_1} + N_{A_0}}\left(\frac{\theta_{n1} - \dfrac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}}}{\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}}}\right),$$

$$\theta_{p0} = \frac{N_{p0} - N_{A_1}}{N_{A_1} + N_{A_0}}\left(\frac{\theta_{n1} - \dfrac{N_{n1} - N_{p0}}{N_{A_1} - N_{p0}}}{\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}}}\right),$$

$$\theta_{a0} = \frac{N_{A_1} + N_{a0}}{N_{A_1} + N_{A_0}}\left(\frac{\theta_{n1} - \dfrac{N_{n1} + N_{a0}}{N_{A_1} + N_{a0}}}{\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}}}\right). \tag{3}$$

Substituting them to the restriction $\theta_{n1}\theta_{a1}\theta_{p0} = \theta_{p1}\theta_{n0}\theta_{a0}$, we obtain the likelihood equation of θ_{n1}

$$c_0 \theta_{n1} (\theta_{n1} - \frac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}})(\theta_{n1} - \frac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}})((N_{A_1} - N_{p0})\theta_{n1} - (N_{n1} - N_{p0}))$$

$$= (\theta_{n1} - \frac{N_{n1} + N_{p1}}{N_{A_1} + N_{p1}})(\theta_{n1} - \frac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}})(\theta_{n1} - \frac{N_{n1} + N_{a0}}{N_{A_1} + N_{a0}}), \tag{4}$$

where $c_0 = [(N_{n1} + N_{p1})(N_{A_1} + N_{A_0})]/[(N_{A_1} + N_{p1})(N_{A_1} + N_{n0})(N_{A_1} + N_{a0})]$ is a positive

constant. Solving the order 4 equation (4) of a single parameter θ_{n1}, we can find at most four

real solutions. We show in Appendix that at least one of them satisfy the normal restrictions of

probabilities. Then we compute the other $\theta_{s'a}$ by those functions of θ_{n1} in (3). Comparing their

likelihood values, we can find θ^{t+1} with the maximum likelihood value in the $(t+1)$th iteration.

The following theorem shows that the above EM algorithm is correct for finding the MLE of θ

under the restrictions of $\{\theta_{s'a}\}$, and its proof is given in Appendix. Some simulation results of this

EM algorithm will be given in Section 6.

THEOREM 2: *MLEs of $\{\theta_{s'a}\}$ obtained via the above EM algorithm satisfy the likelihood equa-*

tions and the restrictions $\theta_{n1} + \theta_{p1} + \theta_{a1} = \theta_{n0} + \theta_{p0} + \theta_{a0} = 1$ and $\theta_{n1}\theta_{a1}\theta_{p0} = \theta_{p1}\theta_{n0}\theta_{a0}$.

5. Bounds of Causal Effects

As shown in the previous section, only with the IA assumption, we cannot identify the joint

distribution of Z, S^{obs}, S' and A. In this section, we remove the M and NI assumptions and give

an approach for finding the lower and upper bounds of the causal effects of interest. By the IA

assumption $Z \perp\!\!\!\perp (A, S')$, we have

$$E(S(1)|A = a) = P(S(1) = 1|A = a) = P(S^{obs} = 1|A = a, Z = 1)$$

$$= \frac{P(Z = 0)}{P(Z = 1)} P(Z = 1, S^{obs} = 1, A = a) \frac{1}{P(Z = 0, A = a)}, \text{ and}$$

$$E(S(0)|A = a) = P(S(0) = 1|A = a) = P(S^{obs} = 1|A = a, Z = 0)$$

$$= P(Z = 0, S^{obs} = 1, A = a) \frac{1}{P(Z = 0, A = a)}.$$

Thus the average causal effect of Z on S within level $A = a$ is

$$E(S(1) - S(0) \mid A = a)$$

$$= \frac{P(Z=1, S^{obs}=1, A = a)P(Z=0) - P(Z=0, S^{obs}=1, A = a)P(Z=1)}{P(Z = 1)P(Z = 0, A = a)}.$$

At the right hand side of this equality, only $P(Z = 0, A = a)$ is unobserved, and we can see that $E(S(1) - S(0) \mid A = a)$ is a monotonic function with respect to $P(Z = 0, A = a)$. Note that the sign of $E(S(1) - S(0) \mid A = a)$ is decided by the numerator defined as Δ_a, which can be directly estimated from the observed data. In addition, if the monotonicity assumption holds, Δ_a is nonnegative, which implies a consistent nonnegative causal effect. The bounds of $E(S(1) - S(0) \mid A = a)$ can be obtained by using the bounds of $P(Z = 0, A = a)$, which can be expressed as $(F_{00a} + F_{01a})/N$. Since $F_{000} + F_{001} = F_{00+}$, we can focus on the bounds of F_{000}. It is obvious that

$$0 \leqslant F_{000} \leqslant F_{00+}.$$

By the IA assumption $Z \perp\!\!\!\perp A$, we have $P(Z = 0, A = 0)P(Z = 1, A = 1) = P(Z = 0, A = 1)P(Z = 1, A = 0)$. By replacing $P(Z = z, A = a)$ with $(F_{z0a} + F_{z1a})/N$, F_{100} can be represented as a function of F_{000}: $F_{100} = [F_{000}F_{1++} + F_{010}(F_{10+} + F_{111}) - F_{110}(F_{00+} + F_{011})]/F_{0++}$. Thus by the inequality $0 \leqslant F_{100} \leqslant F_{10+}$, we can find another bound of F_{000}:

$$[F_{110}(F_{00+} + F_{011}) - F_{010}(F_{10+} + F_{111})]/F_{1++}$$

$$\leqslant F_{000} \leqslant [(F_{00+} + F_{011})(F_{10+}F_{110}) - F_{010}F_{111}]/F_{1++}.$$

Combining the above two sets of bounds, we can take the common range of them as the bounds of F_{000}: $F_{000}^{L} \leqslant F_{000} \leqslant F_{000}^{U}$, where

$$F_{000}^{L} = \max\{0, [F_{110}(F_{00+} + F_{011}) - F_{010}(F_{10+} + F_{111})]/F_{1++}\} \text{ and}$$

$$F_{000}^{U} = \min\{F_{00+}, [(F_{00+} + F_{011})(F_{10+} + F_{110}) - F_{010}F_{111}]/F_{1++}\}.$$

Since $F_{001} = F_{00+} - F_{000}$, we have $F_{001}^{L} = F_{00+} - F_{000}^{U} \leqslant F_{001} \leqslant F_{00+} - F_{000}^{L} = F_{001}^{U}$. When $\Delta_a \geqslant 0$, the boundary of $E(S(1) - S(0) \mid A = a)$ is $[\ (N\Delta_a)/(P_Z F_{00a}^{U})\ ,\ (N\Delta_a)/(P_Z F_{00a}^{L})\ ]$, and the opposite lower and upper bounds are for $\Delta_a < 0$.

6. Simulation

In this section, we evaluate the estimation approaches presented in Sections 3 and 4 via simulations. We perform the simulations for two cases, one with the SM assumption and the other without the SM assumption. Given the true values of parameters, we draw random samples from the corresponding binomial distribution and estimate the parameters by the EM algorithm. Repeating this process 1000 times, we find the means and mean squared errors (MSE) of estimates. In Tables 5 and 6, means and MSEs (in brackets) of MLEs of parameters are shown for two different sets of true parameters under the SM assumption, while the results for two different sets of true parameters without the SM assumption are given in Tables 7 and 8.

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

Let m denote the missing rate. For the data missing due to death, the missing rate depends on the values of parameters. Since the missing part consists of 'never survivors' and the 'protectable' assigned to the standard treatment, we have the missing rate $m = P(S' = \mathrm{n}) + P(Z = 0)P(S' = \mathrm{p})$. Furthermore, under the SM assumption, the missing rate $m = (1 - p_Z)(p_A(1 - \theta_1) + (1 - p_A)(1 - \theta_0))$; without requiring the SM assumption, the rate $m = p_A\theta_{\mathrm{n}1} + (1 - p_A)\theta_{\mathrm{n}0} + (1 - p_Z)(p_A\theta_{\mathrm{p}1} + (1 - p_A)\theta_{\mathrm{p}0})$. Thus the missing rates are 72.9%, 37.5%, 74.0% and 54.5% for Tables 5 to 8 respectively.

For all scenarios, the MSEs decrease as the sample size increases. Comparing MSEs in Tables 5 and 6, for the same assumptions and the same sample size, the MSEs for the case with a larger missing rate are larger than those for a smaller missing rate. Further the MSEs under the SM assumption are smaller than those under the IA, M and NI assumptions. All of these results accord with our intuition.

7. Results of NSCOT data

The artificial data of NSCOT in Table 9 are generated from the percentages in Table 2 of Frangakis et al. (2007) since the real data are not available. As defined in Section 2, Z, A and S denote the treatment, physical condition and survival status respectively. Here X could be the injury severity. $Z = 1$ denotes that the transport time is less than 10 minutes. In the NSCOT data, there are no deaths for people delivered to the hospital within 10 minutes after a critical event (such as an accident) happened, which is required by the SM assumption. $\theta_1 = P(S(0) = 1 \mid A = 1)$ and $\theta_0 = P(S(0) = 1 \mid A = 0)$ are the livability of standard treatment group under poor and good physical conditions respectively. From the results in Table 10, we can see that $\theta_0 > \theta_1$ means that healthier persons have a larger survival probability. During the critical event, the high injured persons have a greater potential risk of death than the low injured ones no matter which physical condition they have (i.e., $0.4435 < 0.8325$ and $0.8427 < 0.9903$). For the case of low-injury severity $X = 0$, the average causal effects of Z on S (i.e., $E(S(1) - S(0) \mid A = a) = 1 - \theta_a$) are 0.168 for the poor physical condition group $A = 1$ and 0.010 for the good one $A = 0$. For the case of high-injury severity $X = 1$, the results are 0.557 and 0.157 respectively. Furthermore, we define the relative ACE: $RACE = E(S(1) - S(0) \mid A = 1)/E(S(1) - S(0) \mid A = 0) = (1 - \theta_1)/(1 - \theta_0)$, which reflects the relative effect of treatment between different groups. Then we obtain that $RACE = 17.46$ conditional on the high-injury severity $X = 1$ and $RACE = 3.54$ conditional on the low-injury severity $X = 0$. Therefore the treatment is more effective for people in poor health condition (i.e., $RACE > 1$), and the timely succor like transporting to the hospital rapidly is more important when the persons are in the bad situation $X = 1$ (i.e., $17.46 > 3.54$).

[Table 9 about here.]

[Table 10 about here.]

Below to illustrate our estimation without requiring the SM assumption, we suppose that there were some persons who would die even if they were delivered quickly, i.e., there existed 'never

survivors'. For example, there were $F_{10+} = 2$ 'never survivors' in high-injury severity group as shown in Table 11, which does not violate the monotonicity assumption. We have $E(S(1) - S(0) \mid A = a) = P(S(1) = 1 \mid A = a) - P(S(0) = 1 \mid A = a) = P(S' = \mathrm{p} \mid A = a)$. From Table 12, the ACE is 0.160 for the poor physical condition group $A = 1$ and 0.004 for the good one $A = 0$, and $RACE = \theta_{p1}/\theta_{p0} = 39.85$. Without requiring the monotonicity nor no-interaction assumption, we can compute the bounds of ACE for the level $A = a$ with the method presented in Section 5: $E(S(1) - S(0) \mid A = 0) \in (0.0034, 0.0040)$, while $E(S(1) - S(0) \mid A = 1) \in (0.1586, 0.4758)$. This means that the causal effects of Z on S are positive conditional on A, and the treatment is more effective for people in poor health condition (i.e. $0.1586 > 0.0040$).

[Table 11 about here.]

[Table 12 about here.]

References

Angrist, J. D. , Imbens, G. W. , and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with Discussion). *Journal of the American Statistical Association* **91,** 444–472

Frangakis, C. E. , Rubin, D. B. , An, M. , and MacKenzie, E. (2007). Principal Stratification Designs to Estimate Input Data Missing Due to Death (with Discussion). *Biometrics* **63,** 641–662.

MacKenzie, E. J., Rivara, F. P., Jurkovich, G. J., Nathens, A. B., Frey, K. P., Egleston, B. L., Salkever, D. S., and Scharfstein, D. O. (2006). A national evaluation of the effect of trauma-center care of mortality. *New England Journal of Medicine*, **354,** 366–378.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: the Fisher Randomization Test Comment. *Journal of the American Statistical Association* **75,** 591–593.

Xie, Y., and Murphy, S. (2007). Discussion in Principal Stratification Designs to Estimate Input Data Missing Due to Death. *Biometrics* **63,** 641–662.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by 'death'. *Journal of Educational and Behavioral Statistics* **28,** 353–368.

Appendix

Proof of Theorem 1

Below we prove Theorem 1 step by step. First at Step 1, we introduce two parameters x_1 and x_2 to represent all parameters to be identified. Then at Steps 2 and 3, we get two equations of x_1 and x_2 from the IA and NI assumptions. At Step 4, we reduce x_2 to get a quadratic equation of x_1. Finally at Step 5, we show that there is a unique proper solution of x_1. Thus all parameters are identified by the unique x_1.

(1) Define $x_1 = P(A = 1 \mid Z = 1, S^{obs} = 0)$ and $x_2 = P(A = 1 \mid Z = 0, S^{obs} = 0)$, which are not identifiable due to death without any further assumption. From the definitions of parameters, we have the following relationships among the parameters

$$p_Z = P(Z = 1)$$

$$p_A = P(A = 1) = P(A = 1, S^{obs} = 1 \mid Z = 1) + x_1 P(S^{obs} = 0 \mid Z = 1),$$

$$\theta_{n1} = P(S' = (0,0) \mid A = 1) = \frac{x_1 P(S^{obs} = 0 \mid Z = 1)}{p_A},$$

$$\theta_{p1} = P(S' = (0,1) \mid A = 1) = 1 - \theta_{n1} - \theta_{a1},$$

$$\theta_{a1} = P(S' = (1,1) \mid A = 1) = \frac{x_2 P(S^{obs} = 0 \mid Z = 0)}{p_A},$$

$$\theta_{n0} = P(S' = (0,0) \mid A = 0) = \frac{(1 - x_1)P(S^{obs} = 0 \mid Z = 1)}{1 - p_A},$$

$$\theta_{p0} = P(S' = (0,1) \mid A = 0) = 1 - \theta_{n0} - \theta_{a0},$$

$$\theta_{a0} = P(S' = (1,1) \mid A = 0) = \frac{(1 - x_2)P(S^{obs} = 0 \mid Z = 0)}{1 - p_A}.$$

p_Z can always be identifiable from the observed data directly, and other parameters are functions of x_1 and x_2. Therefore we only need to focus on the identifiability of these two probabilities x_1 and x_2.

(2) By the IA assumption, we can have the following relationship between x_2 and x_1:

$$P(A = 1 \mid Z = 1) = P(A = 1 \mid Z = 0)$$

$$\Rightarrow \quad P(A = 1, S^{obs} = 1 \mid Z = 1) + x_1 P(S^{obs} = 0 \mid Z = 1)$$

$$= \quad P(A = 1, S^{obs} = 1 \mid Z = 0) + x_2 P(S^{obs} = 0 \mid Z = 0)$$

$$\Rightarrow \quad x_2 = a + bx_1 \,,$$

where $a = [P(A = 1, S^{obs} = 1 \mid Z = 1) - P(A = 1, S^{obs} = 1 \mid Z = 0)]/P(S^{obs} = 0 \mid Z = 0)$ and $b = P(S^{obs} = 0 \mid Z = 1)/P(S^{obs} = 0 \mid Z = 0)$, both of which are identifiable. According to the relationship of x_1 and x_2, we only need to show that one of them is identifiable. Below we prove that x_1 is identifiable.

(3) By the NI assumption, we have another relationship between x_1 and x_2

$$\frac{x_1(1 - x_2)}{x_2(1 - x_1)} = \frac{P(A = 1, Z = 1, S^{obs} = 1)P(A = 0, Z = 0, S^{obs} = 1)}{P(A = 1, Z = 0, S^{obs} = 1)P(A = 0, Z = 1, S^{obs} = 1)}.$$

Since we can observe complete data for every survivor, the right hand side of the above equation for the survivals $S^{obs} = 1$ is identifiable, denoted by c. Then we have another equation $cx_2(1 - x_1) = x_1(1 - x_2)$.

(4) Replacing x_2 by $a + bx_1$ in the last formula, we obtain a quadratic equation of x_1

$$z_1 + z_2 x_1 + z_3 x_1^2 = 0 \,, \tag{$*$}$$

where the coefficients $z_1 = ca$, $z_2 = cb - ca + a - 1$ and $z_3 = b - cb$, which are all identifiable. If this equation has a unique root in the interval $(0, 1)$ for the probability x_1, then x_1 is identifiable.

(5) By Table 4, we can use those parameters to express the coefficients in equation $(*)$ as

$$a = \frac{p_A \theta_{p1}}{p_A(\theta_{n1} + \theta_{p1}) + (1 - p_A)(\theta_{n0} + \theta_{p0})} \,,$$

$$b = \frac{p_A \theta_{n1} + (1 - p_A)\theta_{n0}}{p_A(\theta_{n1} + \theta_{p1}) + (1 - p_A)(\theta_{n0} + \theta_{p0})} \,,$$

$$c = \frac{\theta_{a0}(\theta_{p1} + \theta_{a1})}{\theta_{a1}(\theta_{p0} + \theta_{a0})},$$

$$\frac{z_1}{z_3} = \frac{p_A \theta_{n1}(1 - \theta_{n1})}{[p_A \theta_{n1} + (1 - p_A)\theta_{n0}](\theta_{n0} - \theta_{n1})},$$

$$\frac{z_2}{z_3} = \frac{p_A \theta_{n1}}{p_A \theta_{n1} + (1 - p_A)\theta_{n0}} + \frac{1 - \theta_{n1}}{\theta_{n0} - \theta_{n1}}.$$

From the last two equalities, we can find the following two real roots of equation $(*)$

$$\frac{p_A \theta_{n1}}{p_A \theta_{n1} + (1 - p_A)\theta_{n0}} \quad \text{and} \quad \frac{1 - \theta_{n1}}{\theta_{n0} - \theta_{n1}}.$$

Note that the second root $(1 - \theta_{n1})/(\theta_{n0} - \theta_{n1})$ is outside the interval $(0, 1)$ because all $\theta_{s'a}$ as probabilities should be in $(0, 1)$. It is easily to see that the other root belongs to this range. Furthermore, we have

$$x_1 = P(A = 1 \mid Z = 1, S^{obs} = 0) = P(A = 1 \mid Z = 1, S' = \text{n})$$

$$= \frac{P(A = 1, S' = \text{n})}{P(S' = \text{n})}$$

$$= \frac{P(A = 1)P(S' = (0, 0) \mid A = 1)}{P(A = 1)P(S' = (0, 0)|A = 1) + P(A = 0)P(S' = (0, 0)|A = 0)}.$$

Thus we obtain

$$x_1 = \frac{p_A \theta_{n1}}{p_A \theta_{n1} + (1 - p_A)\theta_{n0}}.$$

When the sample size is large enough, all the straight estimated values of coefficients from the observational data will converge to the theoretic true value, and we can identify x_1 by solving equation $(*)$ to get the unique root which is within the interval $(0, 1)$.

The above proof does not only prove the identifiability but also gives a method for estimating these parameters. First, we estimate the probabilities by the sample means, next calculate the coefficients (z_1, z_2, z_3) in equation $(*)$, then solve this equation to obtain the estimate of x_1, and finally compute estimates of other parameters with the relations between x_1 and these parameters.

Proof of Theorem 2

The key point of this proof is to show that the M-step of this EM algorithm is correct for finding θ which maximizes $L(\theta)$ under those restrictions in every iteration when $N_{s'a} \neq 0$.

Since every conditional probability $\theta_{s'a}$ should be in the interval $[0, 1]$ and every other $\theta_{s'a}$ can be

expressed as a function of θ_{n1} as shown in Section 4.3, we have the following inequalities

$$0 \leqslant \theta_{n1} \leqslant 1,$$

$$0 \leqslant \theta_{p1} \leqslant 1 \Leftrightarrow \frac{N_{n1} - N_{a1}}{N_{A_1} + N_{p1}} \leqslant \theta_{n1} \leqslant \frac{N_{n1} + N_{p1}}{N_{A_1} + N_{p1}},$$

$$0 \leqslant \theta_{a1} \leqslant 1 \Leftrightarrow \frac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}} \leqslant \theta_{n1} \leqslant 1,$$

$$0 \leqslant \theta_{n0} \leqslant 1 \Leftrightarrow \theta_{n1} \leqslant \frac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}},$$

$$0 \leqslant \theta_{p0} \leqslant 1 \Leftrightarrow \begin{cases} \theta_{n1} \leqslant \dfrac{N_{n1} + (N_{n0} + N_{a0})/2}{N_{A_1} + (N_{n0} + N_{a0})/2} & \text{if } N_{p0} > N_{A_1}, \\[2mm] \dfrac{N_{n1} - N_{p0}}{N_{A_1} - N_{p0}} \leqslant \theta_{n1} \leqslant \dfrac{N_{n1} + (N_{n0} + N_{a0})/2}{N_{A_1} + (N_{n0} + N_{a0})/2} & \text{if } N_{p0} < N_{A_1}, \\[2mm] \theta_{n1} \leqslant \dfrac{2N_{n1} + N_{A_0} - N_{A_1}}{N_{A_1} + N_{A_0}} & \text{if } N_{p0} = N_{A_1}, \end{cases}$$

$$0 \leqslant \theta_{a0} \leqslant 1 \Leftrightarrow \theta_{n1} \leqslant \frac{N_{n1} + N_{a0}}{N_{A_1} + N_{a0}}.$$

The above boundaries of θ_{n1} can be transformed to the following restrictions

(1) for the case of $N_{p0} < N_{A_1}$,

$$\max\left\{0, \frac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}}, \frac{N_{n1} - N_{p0}}{N_{A_1} - N_{p0}}\right\} \leqslant \theta_{n1}$$

$$\leqslant \min\left\{\frac{N_{n1} + N_{p1}}{N_{A_1} + N_{p1}}, \frac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}}, \frac{N_{n1} + N_{a0}}{N_{A_1} + N_{a0}}\right\},$$

(2) for the case of $N_{p0} \geqslant N_{A_1}$,

$$\max\left\{0, \frac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}}\right\} \leqslant \theta_{n1}$$

$$\leqslant \min\left\{\frac{N_{n1} + N_{p1}}{N_{A_1} + N_{p1}}, \frac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}}, \frac{N_{n1} + N_{a0}}{N_{A_1} + N_{a0}}\right\}.$$

Let $f(\theta_{n1})$ and $g(\theta_{n1})$ denote the left and right hand sides of the likelihood equation (4), where

$f(\theta_{n1})$ and $g(\theta_{n1})$ are a quartic polynomial and a cubic one respectively. Note that every element in

the above lower bound of θ_{n1} is the root of equation $f(\theta_{n1}) = 0$ and that the three elements in the

above upper bound of θ_{n1} are just the three roots of equation $g(\theta_{n1}) = 0$. Let ϕ_1 and ϕ_2 denote the

lower and upper bound of θ_{n1}. Thus we have that ϕ_1 is one root of equation $f(\theta_{n1}) = 0$ and that ϕ_2

is the smallest root of equation $g(\theta_{n1}) = 0$. Since $\phi_1 \leqslant \phi_2$ and the coefficient of the highest order term in $g(\theta_{n1})$ is 1, we always have $g(\phi_1) \leqslant 0$. Below we prove that there must exist at least one root in the required restrictions to satisfy $f(\theta_{n1}) = g(\theta_{n1})$.

(1) When $N_{p0} < N_{A_1}$, we have $f(\theta_{n1}) = c_1 \theta_{n1}(\theta_{n1} - \dfrac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}})(\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}})(\theta_{n1} - \dfrac{N_{n1} - N_{p0}}{N_{A_1} - N_{p0}})$,

where c_1 is a positive constant. By the definitions of ϕ_1 and ϕ_2, we have $\dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}} > \dfrac{N_{n1} + N_{n0}}{N_{A_1} + N_{n0}} \geqslant$

$\phi_2 \geqslant \phi_1$, which implies that ϕ_1 is the second largest root of $f(\theta_{n1}) = 0$ and that ϕ_2 is smaller

than the largest root $\dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}}$. Therefore we have $f(\phi_2) \leqslant 0$.

(2) When $N_{p0} > N_{A_1}(> N_{n1})$, we have $f(\theta_{n1}) = c_2 \theta_{n1}(\theta_{n1} - \dfrac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}})(\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}})(\theta_{n1} -$

$\dfrac{N_{p0} - N_{n1}}{N_{p0} - N_{A_1}})$, where c_2 is a negative constant. Since $\dfrac{N_{p0} - N_{n1}}{N_{p0} - N_{A_1}} > 1 > \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}} > \phi_2 \geqslant \phi_1$, ϕ_1

is the third largest root of $f(\theta_{n1}) = 0$, and ϕ_2 is smaller than the second largest root. This also

means $f(\phi_2) \leqslant 0$.

(3) When $N_{p0} = N_{A_1}$, we have $f(\theta_{n1}) = c_3 \theta_{n1}(\theta_{n1} - \dfrac{N_{n1} - N_{a1}}{N_{n1} + N_{p1}})(\theta_{n1} - \dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}})$, where c_3 is a

positive constant. We still have $f(\phi_2) \leqslant 0$ because ϕ_1 is the second largest root of $f(\theta_{n1}) = 0$

and ϕ_2 is smaller than the largest root $\dfrac{N_{n1} + N_{A_0}}{N_{A_1} + N_{A_0}}$.

Therefore $f(\phi_2) \leqslant 0$ holds for all of the above three cases. Let $d(\theta_{n1}) = f(\theta_{n1}) - g(\theta_{n1})$. We have

$d(\phi_1) = 0 - g(\phi_1) \geqslant 0$ and $d(\phi_2) = f(\phi_2) - 0 \leqslant 0$, and thus there must exist at least one solution

to equation $d(\theta_{n1}) = 0$ in the interval $[\phi_1, \phi_2]$ from the property of continuous function $d(\theta_{n1})$. In

addition, there are at most four different real solutions to $f(\theta_{n1}) = g(\theta_{n1})$ under the restrictions.

Thus we can solve this equation and then compare the likelihood functions to find the maximum

one.

Table 1
Frequencies with missing due to death with the SM assumption

Z	S^{obs}	$A = 0$	$A = 1$	Subtotal
0	0	$F_{000} = ?$	$F_{001} = ?$	F_{00+}
	1	F_{010}	F_{011}	F_{01+}
1	0	$F_{100} = 0$	$F_{101} = 0$	$F_{10+} = 0$
	1	F_{110}	F_{111}	F_{11+}
Total				N

Table 2
Potentially completely frequencies under the SM assumption

Z	A	S'	Potential $f_{zas_0s_1}$	Probability
0	0	(0,1)	f_{0001}	$(1 - p_Z)(1 - p_A)(1 - \theta_0)$
		(1,1)	F_{010}	$(1 - p_Z)(1 - p_A)\theta_0$
	1	(0,1)	$F_{00+} - f_{0001}$	$(1 - p_Z)p_A(1 - \theta_1)$
		(1,1)	F_{011}	$(1 - p_Z)p_A\theta_1$
1	0	(0,1)	$F_{110} - f_{1011}$	$p_Z(1 - p_A)(1 - \theta_0)$
		(1,1)	f_{1011}	$p_Z(1 - p_A)\theta_0$
	1	(0,1)	$F_{111} - f_{1111}$	$p_Z p_A(1 - \theta_1)$
		(1,1)	f_{1111}	$p_Z p_A\theta_1$

Table 3
Frequencies with missing due to death without the SM assumption

Z	S^{obs}	$A = 0$	$A = 1$	Subtotal
0	0	$F_{000} = ?$	$F_{001} = ?$	F_{00+}
	1	F_{010}	F_{011}	F_{01+}
1	0	$F_{100} = ?$	$F_{101} = ?$	F_{10+}
	1	F_{110}	F_{111}	F_{11+}
Total				N

Table 4
Potentially complete frequencies under the M assumption

Z	A	S'	Potential $f_{zas_0s_1}$	Probability
0	0	(0,0)	f_{0000}	$(1-p_Z)(1-p_A)\theta_{n0}$
		(0,1)	$F_{000} - f_{0000}$	$(1-p_Z)(1-p_A)\theta_{p0}$
		(1,1)	F_{010}	$(1-p_Z)(1-p_A)(1-\theta_{n0}-\theta_{p0})$
	1	(0,0)	f_{0100}	$(1-p_Z)p_A\theta_{n1}$
		(0,1)	$F_{001} - f_{0100}$	$(1-p_Z)p_A\theta_{p1}$
		(1,1)	F_{011}	$(1-p_Z)p_A(1-\theta_{n1}-\theta_{p1})$
1	0	(0,0)	F_{100}	$p_Z(1-p_A)\theta_{n0}$
		(0,1)	f_{1001}	$p_Z(1-p_A)\theta_{p0}$
		(1,1)	$F_{110} - f_{1001}$	$p_Z(1-p_A)(1-\theta_{n0}-\theta_{p0})$
	1	(0,0)	F_{101}	$p_Zp_A\theta_{n1}$
		(0,1)	f_{1101}	$p_Zp_A\theta_{p1}$
		(1,1)	$F_{111} - f_{1101}$	$p_Zp_A(1-\theta_{n1}-\theta_{p1})$

Table 5
Results 1 under the IA and SM assumptions (m = 72.9%)

Sample Size	$p_Z = 0.10$	$p_A = 0.10$	$\theta_1 = 0.10$	$\theta_0 = 0.20$
50	0.101	0.105	0.283	0.204
	(0.0009)	(0.0097)	(0.1818)	(0.0030)
100	0.100	0.098	0.209	0.201
	(0.0005)	(0.0041)	(0.0985)	(0.0012)
500	0.100	0.100	0.126	0.200
	(0.0002)	(0.0018)	(0.0141)	(0.0005)
1000	0.100	0.100	0.105	0.200
	(0.0001)	(0.0005)	(0.0013)	(0.0001)

Table 6
Results 2 under the IA and SM assumptions (m = 37.5%)

Sample Size	$p_Z = 0.50$	$p_A = 0.50$	$\theta_1 = 0.20$	$\theta_0 = 0.30$
50	0.498	0.498	0.197	0.307
	(0.0023)	(0.0051)	(0.0085)	(0.0134)
100	0.500	0.499	0.202	0.302
	(0.0013)	(0.0024)	(0.0040)	(0.0059)
500	0.500	0.501	0.201	0.302
	(0.0005)	(0.0010)	(0.0016)	(0.0027)
1000	0.500	0.500	0.200	0.299
	(0.0003)	(0.0005)	(0.0008)	(0.0011)

Table 7
Results 1 under the IA, M and NI assumptions (m = 74.0%)

Sample Size	$p_Z = 0.10$	$p_A = 0.10$	$\theta_{n1} = 0.20$	$\theta_{p1} = 0.70$	$\theta_{a1} = 0.10$	$\theta_{n0} = 0.10$	$\theta_{p0} = 0.70$	$\theta_{a0} = 0.20$
50	0.101	0.112	0.252	0.458	0.289	0.085	0.709	0.206
	(0.0009)	(0.0111)	(0.1186)	(0.2081)	(0.2071)	(0.0092)	(0.0112)	(0.0030)
100	0.101	0.111	0.207	0.538	0.255	0.090	0.704	0.206
	(0.0004)	(0.0071)	(0.0504)	(0.1143)	(0.1459)	(0.0041)	(0.0050)	(0.0017)
500	0.100	0.109	0.198	0.651	0.151	0.093	0.703	0.204
	(0.0002)	(0.0035)	(0.0191)	(0.0229)	(0.0307)	(0.0020)	(0.0021)	(0.0006)
1000	0.100	0.103	0.199	0.690	0.111	0.097	0.702	0.201
	(0.0001)	(0.0010)	(0.0056)	(0.0021)	(0.0028)	(0.0005)	(0.0005)	(0.0002)

Table 8
Results 2 under the IA, M and NI assumptions (m = 54.5%)

Sample Size	$p_Z = 0.50$	$p_A = 0.50$	$\theta_{n1} = 0.40$	$\theta_{p1} = 0.40$	$\theta_{a1} = 0.20$	$\theta_{n0} = 0.28$	$\theta_{p0} = 0.42$	$\theta_{a0} = 0.30$
50	0.499	0.491	0.354	0.406	0.240	0.261	0.400	0.339
	(0.0026)	(0.0218)	(0.0241)	(0.0112)	(0.0279)	(0.0214)	(0.0129)	(0.0367)
100	0.500	0.494	0.371	0.411	0.219	0.268	0.416	0.316
	(0.0013)	(0.0125)	(0.0131)	(0.0044)	(0.0120)	(0.0126)	(0.0049)	(0.0166)
500	0.500	0.495	0.386	0.405	0.209	0.277	0.418	0.305
	(0.0005)	(0.0060)	(0.0056)	(0.0017)	(0.0042)	(0.0061)	(0.0018)	(0.0060)
1000	0.500	0.497	0.395	0.401	0.204	0.281	0.419	0.301
	(0.0001)	(0.0017)	(0.0015)	(0.0005)	(0.0010)	(0.0019)	(0.0005)	(0.0017)

Table 9
The artificial NSCOT data under the SM assumption

Z	S^{obs}	$A = 0$	$A = 1$	Subtotal
		X = low-injury severity		
0	0	?	?	17
	1	257	72	329
1	0	0	0	0
	1	6	2	8
Total				354
		X = high-injury severity		
0	0	?	?	24
	1	95	5	100
1	0	0	0	0
	1	10	1	11
Total				135

Table 10

Results for the first NSCOT data set

	p_Z	p_A	θ_1	θ_0
X = low-injury severity	0.0226	0.2500	0.8325	0.9903
X = high-injury severity	0.0815	0.0909	0.4435	0.8427

Table 11
The artificial NSCOT data under the IA, M and NI assumptions

Z	S^{obs}	A = 0	A = 1	Subtotal
		X = high-injury severity		
0	0	?	?	24
	1	95	5	100
1	0	?	?	2
	1	10	1	11
Total				137

Table 12

Results for the second NSCOT data set

p_Z	p_A	θ_{n1}	θ_{p1}	θ_{n0}	θ_{p0}
0.09489	0.22837	0.66326	0.16017	0.00303	0.00411