



## Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates

Christina A. Holcroft<sup>a,\*</sup>,<sup>1</sup>, Andrea Rotnitzky<sup>b,2</sup>, James M. Robins<sup>b,3</sup>

<sup>a</sup>Department of Work Environment, University of Massachusetts Lowell, 1 University Avenue, Lowell, MA 01854, U.S.A.

<sup>b</sup>Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

Received 20 May 1996; received in revised form 11 March 1997

---

### Abstract

Often the variables in a regression model are difficult or expensive to obtain so auxiliary variables are collected in a preliminary step of a study and the model variables are measured at later stages on only a subsample of the study participants called the validation sample. We consider a study in which at the first stage some variables, throughout called auxiliaries, are collected; at the second stage the true outcome is measured on a subsample of the first-stage sample, and at the third stage the true covariates are collected on a subset of the second-stage sample. In order to increase efficiency, the probabilities of selection into the second and third-stage samples are allowed to depend on the data observed at the previous stages. In this paper we describe a class of inverse-probability-of-selection-weighted semiparametric estimators for the parameters of the model for the conditional mean of the outcomes given the covariates. We assume that a subject's probability of being sampled at subsequent stages is bounded away from zero and depends only on the subject's data collected at the previous sampling stages. We show that the asymptotic variance of the optimal estimator in our class is equal to the semiparametric variance bound for the model. Since the optimal estimator depends on unknown population parameters it is not available for data analysis. We therefore propose an adaptive estimation procedure for locally efficient inferences. A simulation study is carried out to study the finite sample properties of the proposed estimators. © 1997 Elsevier Science B.V.

*Keywords:* Measurement error; Missing at random; Auxiliary variable

---

---

\* Corresponding author.

<sup>1</sup> Research supported by National Institutes of Health Grants 5T32 CA09337-12&13&14 and 1-R29-GM48704-01A1.

<sup>2</sup> Research supported in part by National Institutes of Health Grant 1-R29-GM48704-01A1.

<sup>3</sup> Research supported in part by National Institutes of Health Grants 2 P30 ES00002, R01AI32475, and R01-ES03405, K04-ES00180.

## 1. Introduction

In epidemiological studies, it is often of interest to estimate the parameters  $\alpha_0$  indexing the conditional mean of the outcome variable  $Y$  given a set of covariates  $(X, V)$ . Some or all of the variables  $Y$  and  $X$  may be difficult or expensive to obtain so a vector of auxiliary variables  $Z$  may be collected in a preliminary step of the study. Auxiliary variables may consist of the mismeasured variables of interest  $Y$  and  $X$  or may be other variables that are associated with the outcomes  $Y$  and/or covariates  $X$  of interest. These auxiliary variables can be used to determine which subjects should have the accurate measures of  $Y$  and  $X$  taken in order to maximize efficiency under cost or sample size limitations.

In this paper we describe a class of estimators for the parameters  $\alpha_0$  of the model for the conditional mean of the outcome  $Y$  given the covariates  $(X, V)$  from multistage studies in which accurate measures of  $Y$  and the subset  $X$  of the covariates  $(X, V)$  are measured after measuring the covariates  $V$  and the vector of auxiliary variables  $Z$ . Our methods assume that a subject's probability of being sampled to have his/her true outcome and covariates measured at subsequent stages is non-zero and depends only on the observed data. In particular, we consider a sequential study in which at the first stage the auxiliary variables  $Z$  and the model covariates  $V$  are measured, at the second stage the outcome  $Y$  is measured on a subsample of the first stage sample, and at the third stage the remaining covariates  $X$  are measured on a subset of the second-stage sample.

Our sequential design would offer cost benefits in settings in which error-prone measures of the outcome and covariates are cheaper to obtain than the true outcome which in turn is cheaper to measure than the true covariates. For example, consider a study of the respiratory health effects of indoor air pollution. Relatively inexpensive preliminary information on respiratory health status and indoor air pollution can be obtained from questionnaire-based self-reports of asthma/wheeze episodes and the presence of gas stoves in the homes of the study participants. More expensive but accurate assessment of respiratory health status can be obtained from hospital or physician's records or from the results of forced expiratory exams (FEV, FVC). Furthermore, accurate indoor air pollution assessment requires installation of measuring equipment in the study participant's homes, a costly technique that can often be carried out in only a small subsample of the study cohort.

Inferences about  $\alpha_0$  are usually carried out by specifying a parametric mode for the joint distribution of the outcomes, covariates and auxiliary variables, and then estimating  $\alpha_0$  by maximum likelihood. These methods, however, can be very non-robust to the assumed parametric models for the marginal law of the covariates and the conditional law of the auxiliaries given the outcomes and covariates, which are not of scientific interest. In addition, with non-Gaussian data, fully parametric methods are typically computationally complex because they require numerical approximations to integrals that do not have analytical expressions. In contrast, in this paper we describe semiparametric estimators of  $\alpha_0$  that are numerically simple and are

consistent and asymptotically normal without requiring specification of the law of the covariates or of the conditional law of the auxiliaries given the outcomes and covariates.

Extensive work has been done on semiparametric methods for estimating  $\alpha_0$  when only the covariates are missing. Pepe and Fleming (1991) and Carroll and Wand (1991) consider the case in which the covariate vector  $V$  includes the auxiliary variables  $Z$ . They assume that  $f(Y|X, V)$  is known up to the parameter  $\alpha_0$  and they let  $f(X|V)$  be unrestricted. These authors estimate  $\alpha_0$  with the value of  $\alpha$  that maximizes the so-called 'estimated likelihood' in which the contribution of a unit  $i$  with missing  $X_i$  is equal to  $\int f(Y_i|x, V_i; \alpha) d\hat{F}(x|V_i)$  where  $\hat{F}(x|v)$  is an estimate of  $F(x|v)$ , the cumulative conditional distribution function of  $X$  given  $V$  calculated from the validation sample data. Pepe and Fleming (1991) assume that  $V$  is discrete and estimate  $F(x|V_i)$  with the empirical conditional distribution. Carroll and Wand (1991) allow for continuous  $V_i$  and use a kernel estimator of  $F(x|v)$ . Both Pepe and Fleming (1991) and Carroll and Wand (1991) estimators require that the probabilities of selection of subjects having their  $X$  values ascertained do not depend on their values of  $Y$  or  $X$ . Furthermore, as noted by Robins et al. (1994), these estimators can be inefficient. Horvitz and Thompson (1952), Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), Kalbfleisch and Lawless (1988), Breslow and Cain (1988), Imbens (1992), Flanders and Greenland (1991), and Zhao and Lipsitz (1992) proposed semiparametric estimators of  $\alpha_0$  when  $Y$  is Bernoulli,  $Z$  is discrete and the probability that  $X$  is missing can depend on  $Y$ ,  $Z$  and  $V$ . Recently, Reilly and Pepe (1995) proposed a so called 'mean score' method that extends the estimator of Flanders and Greenland (1991) to discrete, but not necessarily binary,  $Y$ . Robins et al. (1994) proposed a class of estimators that allows for non-discrete variables. They showed that their class of estimators constitutes essentially all regular asymptotically linear (RAL) estimators of  $\alpha_0$  since any RAL estimator of  $\alpha_0$  is asymptotically equivalent to an estimator in their class. In particular, they showed that, with the exception of the computationally difficult estimator of Cosslett (1981), the previously developed semiparametric estimators are asymptotically equivalent to inefficient estimators in their class. In addition, they showed that their class contained a member whose asymptotic variance attained the semiparametric variance bound in the semiparametric model defined solely by restrictions on the conditional mean of  $Y$  given  $X$  and  $V$ , when only the covariates  $X$  are missing and the selection probabilities are positive and depend on  $Y$ ,  $V$  and  $Z$  but not on  $X$ .

Recent research has also focused on the problem of estimating  $\alpha_0$  when  $Y$  is missing for a subset of the study participants but  $X$  is always observed. Pepe (1992) proposed an 'estimated maximum likelihood' approach similar to that of Pepe and Fleming (1991). Under Pepe's (1992) approach the likelihood for  $\alpha_0$  is maximized using a parametric model for  $f(Y|X, V)$  and a non-parametric estimator of  $f(Z|Y, X, V)$  calculated from the subsample of subjects with  $Y$  observed (when only the outcomes  $Y$  are missing, the marginal law of  $(X, V)$  is ancillary and therefore need not be estimated). Pepe's (1992) estimator of  $\alpha_0$  is consistent only when the probability that

$Y$  is missing is independent of the auxiliaries  $Z$ . Pepe et al. (1994) proposed a mean score method similar to that of Reilly and Pepe (1995) that allows for the missingness probabilities to depend on  $Z$  and that does not require full specification of the law  $f(Y|X, V)$ . Rotnitzky and Robins (1995b) described a class of estimators that are consistent and asymptotically normal for estimating  $\alpha_0$  when the law  $f(Y|X, V)$  is parametrically modeled. Their approach does not require assumptions on the law  $f(Z|Y, X, V)$ . Their class contains a member whose asymptotic variance attains the semiparametric variance bound for estimating  $\alpha_0$  in their model. These authors show that both Pepe's (1992) and Pepe et al.'s (1994) estimators are asymptotically equivalent to inefficient estimators in their class. Rotnitzky and Robins (1995a) further relaxed the parametric assumptions on  $f(Y|X, V)$  and proposed a class of estimators that are consistent and asymptotically normal for  $\alpha_0$  in a model that imposes parametric restrictions solely on the conditional mean of  $Y$  given  $X$  and  $V$ . They further showed that a member in their class is semiparametric efficient.

Although extensive research has been done about the cases in which one of  $X$  or  $Y$  are measured on only a subsample of the study cohort, there seems to be a lack of work that addresses the estimation of  $\alpha_0$  when both  $X$  and  $Y$  are missing on a subset of the study participants. The goal of this paper is to provide methods for estimating  $\alpha_0$  under this scenario. In particular, we shall extend the work of Robins et al. (1994) to include studies in which both outcomes and covariates are not observed at the first stage.

The paper is outlined as follows. Our semiparametric model is defined in Section 2. In Section 3 we present a class of estimating equations and discuss the asymptotic properties of its solutions. In Section 4 we show that the asymptotic variance of the optimal solution in our class attains the semiparametric variance bound for  $\alpha_0$  in our model. In Sections 2–4 we assume that the data are missing by design. In Section 5 we extend our methods to allow for data missing by happenstance. Since the optimal solution in our class depends on the unknown full-data distribution it is not available for data analysis. Thus, in Section 6 we describe an adaptive estimation procedure for locally efficient estimation at a parametric submodel. Section 7 shows the results of a simulation study of the finite sample properties of our estimators. We conclude with some final remarks in Section 8.

## 2. The problem

Let  $Y_i$ ,  $i = 1, \dots, n$ , be a (possibly multivariate) outcome variable which is either discrete or continuous, and let  $(X_i^T, V_i^T)^T$  be the associated vector of covariates. Here and throughout,  $T$ , when used as a superscript denotes matrix transposition. We assume that the conditional mean of  $Y_i$  given  $X_i$  and  $V_i$  is known up to a finite vector of parameters, that is,

$$E(Y_i|X_i, V_i) = g(X_i, V_i; \alpha_0) \quad (1)$$

where, for each  $\alpha$ ,  $g(\cdot, \cdot; \alpha)$  is a known smooth  $t \times 1$  vector function of the same dimension as  $Y_i$ , and  $\alpha_0$  is a  $q \times 1$  vector of unknown parameters. Our goal is to estimate  $\alpha_0$  when  $Y_i$  and  $X_i$  are not measured in all study subjects, but instead a (possibly vector) auxiliary variable  $Z_i$  is measured in all study participants. Specifically, we assume that  $(V_i^T, Z_i^T)$  are always observed and that  $(Y_i^T, X_i^T)$  are observed in a subsample of the study cohort called the validation sample.

In many applications  $Z_i$  is just an imperfect measurement of an outcome  $Y_i$  that is difficult or expensive to obtain. For example,  $Y_i$  may represent body lead burden as measured by bone lead levels. Bone lead levels are measured by a technique, called K-XRF, that is lengthy and requires exposure to X-rays. Instead,  $Z_i$  may represent lead measurements from blood samples which are easier and cheaper to obtain. When  $Z_i$  is just  $Y_i$  measured with noise, it is often reasonable to assume that

$$Z_i = Y_i + \varepsilon_i \quad \text{and} \quad E(\varepsilon_i | Y_i, X_i, V_i) = 0. \quad (2)$$

Eq. (2), however, may fail to hold in settings in which the error in  $Z_i$  is associated with the values of the covariates  $X_i$  and  $V_i$  and the outcome  $Y_i$ , that is,  $E(\varepsilon_i | Y_i, X_i, V_i)$  is a function of  $Y_i$ ,  $X_i$  and  $V_i$ . For example, in the air pollution study described in the introduction, (2) will not hold if sick subjects from contaminated homes are more likely to report wheeze episodes than sick subjects from clean homes. Throughout, we will not assume that (2) holds.

In this paper we consider the problem of making inferences about  $\alpha_0$  when the data are collected from a three-stage study design. Specifically, at the first stage, a random sample of  $(V_i^T, Z_i^T)$  is taken. At stage 2, the outcome  $Y_i$  is sampled from the first-stage sample with probability that may depend on  $V_i$  and  $Z_i$ . Finally, at the third stage, the covariates  $X_i$  are sampled among subjects selected at the second stage, with probability that may depend of their values of  $V_i$ ,  $Z_i$  and  $Y_i$ . Formally, if  $\Delta_{1i}$  and  $\Delta_{2i}$  are the indicator variables of selection into the second and third stages respectively, i.e.  $\Delta_{1i} = 1$  if  $Y_i$  is observed and 0 otherwise and  $\Delta_{2i} = 1$  if  $X_i$  is observed and 0 otherwise, the three-stage design specifies that

$$\Pr(\Delta_{1i} = 1 | Y_i, X_i, V_i, Z_i) = \Pr(\Delta_{1i} = 1 | V_i, Z_i) \quad (3)$$

and

$$\Pr(\Delta_{2i} = 1 | \Delta_{1i}, Y_i, X_i, V_i, Z_i) = \Pr(\Delta_{2i} = 1 | \Delta_{1i} = 1, Y_i, V_i, Z_i) \Delta_{1i}. \quad (4)$$

Eqs. (3) and (4) are equivalent to the condition that  $(Y_i^T, X_i^T)$  are missing at random in the sense defined by Rubin (1976). Note that when

$$\Pr(\Delta_{2i} = 1 | \Delta_{1i} = 1, Y_i, V_i, Z_i) = 1, \quad (5)$$

then  $\Delta_{1i} = \Delta_{2i}$  and the three-stage design reduces to a two-stage design in which  $Y_i$  and  $X_i$  are simultaneously ascertained in a subsample selected from the first-stage sample with probability that may depend on  $V_i$  and  $Z_i$ . The estimation methods described in this paper will also be valid for this two-stage design.

Until Section 5 we assume that the probabilities of selection into the validation sample are under the control of the investigator. That is, letting  $\pi_{1i} \equiv \pi_1(V_i, Z_i) \equiv \Pr(\Delta_{1i} = 1 | V_i, Z_i)$  and  $\pi_{2i} \equiv \pi_2(Y_i, V_i, Z_i) \equiv \Pr(\Delta_{2i} = 1 | \Delta_{1i} = 1, Y_i, V_i, Z_i)$  we assume that

$$\pi_1(V_i, Z_i) \text{ and } \pi_2(Y_i, V_i, Z_i) \text{ are known functions.} \quad (6)$$

Furthermore, we assume that

$$\pi_{ji} > \sigma > 0, \quad j = 1, 2 \quad (7)$$

for some  $\sigma$ , so that each subject has probability bounded away from zero of being selected into the validation sample.

If the functions  $\pi_1(V_i, Z_i)$  and  $\pi_2(Y_i, V_i, Z_i)$  are appropriately chosen, the three-stage design defined by Eqs. (3), (4) and (7) will offer efficiency advantages over a three-stage random sample design in which selection is done via simple random sampling, i.e. with  $\pi_1(V_i, Z_i)$  and  $\pi_2(Y_i, V_i, Z_i)$  constant functions (Breslow and Cain, 1988; Flanders and Greenland, 1991; Holcroft et al., 1995; Reilly and Pepe, 1995; Tosteson and Ware, 1990). For example, designs that measure  $Y_i$  more frequently among subjects  $i$  with rare or extreme values of  $Z_i$  and  $V_i$ , and, in turn, measure an expensive covariate  $X_i$  on a smaller subsample that overrepresents the subjects with rare or extreme values of  $Y_i$ ,  $V_i$  and  $Z_i$  will typically result in more precise estimators of  $\alpha_0$  than the estimators calculated from three-stage random sample designs.

For subject  $i$  we call  $L_i = (Y_i^T, X_i^T, V_i^T, Z_i^T)^T$  the full data, and

$$(\Delta_{1i}, \Delta_{2i}, L_{\text{obs},i}^T) \quad (8)$$

the observed data where we define  $L_{\text{obs},i} = L_i$  if  $\Delta_{1i} = \Delta_{2i} = 1$ ,  $L_{\text{obs},i} = (Y_i^T, V_i^T, Z_i^T)^T$  if  $\Delta_{1i} = 1$  and  $\Delta_{2i} = 0$ , and  $L_{\text{obs},i} = (V_i^T, Z_i^T)^T$  if  $\Delta_{1i} = \Delta_{2i} = 0$ . We assume that  $(\Delta_{1i}, \Delta_{2i}, L_i^T)$ ,  $i = 1, \dots, n$ , are independent and identically distributed. A semiparametric model is characterized both by the available data and by restrictions on the joint distribution of the data. Our semiparametric model, throughout called 'three-stage', is defined by restrictions (1), (3), (4), (6), (7) and observed data (8). The 'two-stage' semiparametric model is defined as the 'three-stage' model with the additional restriction (5). The first goal of this paper is to provide a class of estimators of  $\alpha_0$  that are consistent and asymptotically normal under the restrictions of the 'three-stage' model. Since the 'three-stage' semiparametric model does not impose restrictions on the conditional law of  $Z_i$  given  $Y_i$ ,  $X_i$  and  $V_i$ , then, in particular, our estimators will be consistent whether or not Eq. (2) is true. The second goal is to show that in the 'three-stage' model, the asymptotic variance of a member in our class coincides with the semiparametric variance bound for all regular estimators of  $\alpha_0$  in the sense defined by Begun et al. (1983). As we will show, this member is not available for data analysis since it depends on the true

law generating the data. Thus, our third goal is to provide an adaptive procedure for nearly efficient estimation.

### 3. A class of estimators

To motivate our estimators, suppose first that  $\alpha_0$  was naively estimated from a complete-case (possibly non-linear) least-squares analysis. That is, consider the estimating equations

$$\sum_i \Delta_{2i} h(X_i, V_i) \varepsilon_i(\alpha) = 0, \quad (9)$$

where  $h(X_i, V_i)$  is an arbitrary smooth  $q \times t$  matrix function and  $\varepsilon_i(\alpha) = Y_i - g(X_i, V_i; \alpha)$ . The estimators solving (9) use only data from subjects with  $\Delta_{2i} = 1$ , i.e. subjects with measured values of  $Y_i$ ,  $X_i$  and  $V_i$ . Unfortunately, when selection into the validation sample depends on an auxiliary variable  $Z_i$  that is statistically dependent with the outcome  $Y_i$ , or when selection into the third-stage sample depends on  $Y_i$  itself, the solutions to Eqs. (9) may fail to be consistent estimators of  $\alpha_0$ . This is so because  $\Delta_{2i} h(X_i, V_i) \varepsilon_i(\alpha_0)$  may not necessarily have mean zero since subjects in the validation sample are a biased sample. The solutions of Eq. (9) will be consistent for  $\alpha_0$  if selection into the validation study depends only on the covariate  $V_i$ , i.e.  $\pi_{1i}$  and  $\pi_{2i}$  depend on  $V_i$  only, but they will generally be inefficient. This is so since, as we shall see later, with missing data the auxiliary variables  $Z_i$  provide information about  $\alpha_0$  but Eq. (9) does not use the variables  $Z_i$ .

Robins et al. (1994) developed a general theory for making inferences about the parameters of semiparametric models with missing data. Their theory motivates considering estimators  $\hat{\alpha}(h, \phi)$  defined as the solutions to the estimating equations

$$n^{-1} \sum_{i=1}^n D_i(\alpha; h, \phi_1, \phi_2) = 0 \quad (10)$$

where

$$D_i(\alpha; h, \phi_1, \phi_2) = \frac{\Delta_{2i} h(X_i, V_i) \varepsilon_i(\alpha)}{\pi_{1i} \pi_{2i}} - \frac{(\Delta_{1i} - \pi_{1i})}{\pi_{1i}} \phi_1(V_i, Z_i) - \frac{(\Delta_{2i} - \pi_{2i} \Delta_{1i})}{\pi_{1i} \pi_{2i}} \phi_2(Y_i, V_i, Z_i),$$

and  $\phi_1(V_i, Z_i)$  and  $\phi_2(Y_i, V_i, Z_i)$  are arbitrary  $q \times 1$  smooth vector functions chosen by the investigator. The estimating Eqs. (10) use data, including the auxiliary variables  $Z_i$ , from all subjects, not just those in the validation study. Subjects selected only for the first-stage sample contribute the term

$$-\frac{0 - \pi_{1i}}{\pi_{1i}} \phi_1(V_i, Z_i) = \phi_1(V_i, Z_i)$$

to the estimating Eqs. (10), subjects in the second-stage sample that are not in the third-stage sample contribute the terms

$$-\left(\frac{1 - \pi_{1i}}{\pi_{1i}}\right) \phi_1(V_i, Z_i) + \left(\frac{1}{\pi_{1i}}\right) \phi_2(Y_i, V_i, Z_i)$$

and subjects in the validation study contribute the terms

$$\frac{1}{\pi_{1i}\pi_{2i}} h(X_i, V_i) \varepsilon_i(\alpha) - \frac{(1 - \pi_{1i})}{\pi_{1i}} \phi_1(V_i, Z_i) - \frac{(1 - \pi_{2i})}{\pi_{1i}\pi_{2i}} \phi_2(Y_i, V_i, Z_i).$$

Note that the term  $(1/\pi_{1i}\pi_{2i})h(X_i, V_i)\varepsilon_i(\alpha)$  is equal to the  $i$ th subject's contribution to the estimating equation (9) from a complete-case analysis weighted by the inverse of the subject's probability of being selected into the validation sample. When  $\pi_{2i} = 1$ , i.e. the 'two-stage' model is true, the term corresponding to  $\phi_2$  drops out of the estimating equations.

Theorem 1 below states the asymptotic properties of  $\hat{\alpha}(h, \phi)$  under the following regularity conditions. Define  $H(\gamma) = h(X, V)\varepsilon(\gamma)$  with  $\gamma = \alpha$ , and let  $\gamma$  be the parameter space of  $\gamma$ . We assume

1.  $\gamma$  is compact and  $\gamma_0$  lies in the interior of  $\gamma$ ;
2.  $(L_i, \Delta_{1i}, \Delta_{2i})$ ,  $i = (1, \dots, n)$ , are independently and identically distributed;
3.  $\pi_1(V_i, Z_i) > \sigma > 0$  and  $\pi_2(Y_i, V_i, Z_i) > \sigma > 0$  almost surely for some  $\sigma$ ;
4.  $E[H(\gamma)] \neq 0$  if  $\gamma \neq \gamma_0$ ;
5.  $\text{Var}[H(\gamma_0)]$  is finite and positive definite;
6.  $E[\partial H(\gamma_0)/\partial \gamma^T]$  exists and is invertible;
7. There exists a neighborhood  $N$  of  $\gamma_0$  such that  $E[\sup_{\gamma \in N} \|H(\gamma)\|]$ ,  $E[\sup_{\gamma \in N} \|\partial H(\gamma)/\partial \gamma^T\|]$ , and  $E[\sup_{\gamma \in N} \|H(\gamma)H(\gamma)^T\|]$  are all finite, where  $\|A\| = \{\sum_{i,j} A_{ij}^2\}^{1/2}$  for any matrix  $A$  with elements  $A_{ij}$ ;
8.  $f(L, \Delta_1, \Delta_2; \gamma)$  is a regular parametric model with score  $S_\gamma(\gamma) = \partial \log f(L, \Delta_1, \Delta_2; \gamma)/\partial \gamma$  where  $f(L, \Delta_1, \Delta_2; \gamma)$  is a density that differs from the true density  $f(L, \Delta_1, \Delta_2) = f(L, \Delta_1, \Delta_2; \gamma_0)$  only in that  $\gamma$  replaces  $\gamma_0$ ;
9. For all  $\gamma^*$  in a neighborhood  $N$  of  $\gamma_0$ ,  $E_{\gamma^*}[H(\gamma^*)]$  and  $E_{\gamma^*}[\sup_{\gamma \in N} \|H(\gamma)^T H(\gamma)\|]$  is bounded, where  $E_{\gamma^*}$  refers to expectation with respect to the density  $f(L, \Delta_1, \Delta_2; \gamma^*)$ .

**Theorem 1.** Under the regularity conditions 1–9, if the assumptions of the 'three-stage' model hold, then with probability going to 1, Eq. (10) has a unique solution  $\hat{\alpha}(h, \phi)$ . This solution satisfies (a)  $\hat{\alpha}(h, \phi)$  is consistent for estimating  $\alpha_0$ , and (b) the distribution of  $n^{1/2}\{\hat{\alpha}(h, \phi) - \alpha_0\}$  is asymptotically normal with mean zero and covariance  $I^{-1}(h)\Omega(h, \phi)I^{-1}(h)^T$  that can be consistently estimated by  $\hat{I}^{-1}(h)\hat{\Omega}(h, \phi)\hat{I}^{-1}(h)^T$  where  $I(h) = -E\{h(X_i, V_i)\partial g(X_i, V_i; \alpha_0)/\partial \alpha^T\}$ ,  $\Omega(h, \phi) = E\{D_i(\alpha_0; h, \phi)D_i^T(\alpha_0; h, \phi)\}$ , and both  $\hat{I}(h) \equiv n^{-1}\sum_i \partial D_i(\alpha; h, \phi)/\partial \alpha^T$  and  $\hat{\Omega}(h, \phi) \equiv n^{-1}\sum_i D_i(\alpha; h, \phi)D_i^T(\alpha; h, \phi)$  are evaluated at  $\alpha = \hat{\alpha}(h, \phi)$ .

Outlines of the proofs of the theorems and lemmas in this article are provided in the Appendix.

## 4. Efficiency considerations

### 4.1. Optimal estimator

Our next result states that there is a member in our class of estimators  $\hat{\alpha}(h, \phi)$  whose asymptotic variance attains the semiparametric variance bound for the ‘three-stage’ model.

**Theorem 2.** *Under the conditions of Theorem 1, there exist unique functions  $h_{\text{eff}}(X_i, V_i)$  and  $\phi_{\text{eff}}(Y_i, V_i, Z_i)^T = (\phi_{1,\text{eff}}(V_i, Z_i)^T, \phi_{2,\text{eff}}(Y_i, V_i, Z_i)^T)$  such that  $\Omega^{-1}(h_{\text{eff}}, \phi_{\text{eff}})$  equals the semiparametric variance bound in the sense of Begun et al. (1983) for the ‘three-stage’ model. Furthermore,  $I(h_{\text{eff}}) = \Omega(h_{\text{eff}}, \phi_{\text{eff}})$  so that the asymptotic variance of  $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$  attains the bound.*

Theorem 2 follows from the general theory of Robins et al. (1994) on efficient estimation in semiparametric models with data missing at random. Theorem 2 says that there exists no estimator that is locally uniformly asymptotically normal and unbiased for all distributions of the data that satisfy the restrictions of the ‘three-stage’ model whose asymptotic variance is smaller than that of  $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$ .

We now derive  $h_{\text{eff}}$  and  $\phi_{\text{eff}}$ . We first start by determining the functions  $\phi_j^h, j = 1, 2$ , that minimize the asymptotic variance of  $\hat{\alpha}(h, \phi)$  for a fixed choice of  $h$ .

**Lemma 1.** *Under the regularity conditions of Theorem 1, for a fixed  $h(X_i, V_i)$ , the asymptotic variance of  $\hat{\alpha}(h, \phi)$  is minimized at  $\hat{\alpha}(h, \phi^h)$  with  $\phi^h = (\phi_1^h, \phi_2^h)$  where  $\phi_1^h(V_i, Z_i) = E\{h(X_i, V_i)\varepsilon_i(\alpha_0) | V_i, Z_i\}$  and  $\phi_2^h(Y_i, V_i, Z_i) = E\{h(X_i, V_i)\varepsilon_i(\alpha_0) | Y_i, V_i, Z_i\}$ .*

Lemma 1 justifies the inclusion of the terms  $\phi_1$  and  $\phi_2$  in Eqs. (10). Since, in general,  $\phi^h \neq 0$ , Lemma 1 says that we can improve the efficiency of estimators from an inverse-weighted-probability analysis based on only data from the third-stage sample (i.e. an analysis based on Eqs. (10) with  $\phi_1 = \phi_2 = 0$ ) by appropriately choosing the functions  $\phi_1$  and  $\phi_2$ . The optimal function  $\phi^h$  depends on each specific  $h$ . In the following theorem we derive the optimal function  $h_{\text{eff}}$ .

**Theorem 3.**  $h_{\text{eff}}$  is the solution to the functional (integral) equation

$$h_{\text{eff}}(X_i, V_i) = \left\{ \frac{\partial g(X_i, V_i; \alpha_0)}{\partial \alpha} \right\} t(X_i, V_i) + E \left\{ \frac{1 - \pi_{1i}}{\pi_{1i}} \phi_{1,\text{eff}} \varepsilon_i^T | X_i, V_i \right\} t(X_i, V_i) \\ + E \left\{ \frac{1 - \pi_{2i}}{\pi_{1i} \pi_{2i}} \phi_{2,\text{eff}} \varepsilon_i^T | X_i, V_i \right\} t(X_i, V_i), \quad (11)$$

where  $\varepsilon_i = \varepsilon_i(\alpha_0)$ ,  $\phi_{\text{eff}} = \phi^{h_{\text{eff}}}$  and  $t(X_i, V_i) = \{E[\varepsilon_i \varepsilon_i^T / \pi_{1i} \pi_{2i} | X_i, V_i]\}^{-1}$ .

Suppose that data on  $Z_i$  were not collected, then  $\phi_{1,\text{eff}}(V_i) = E[h_{\text{eff}}(X_i, V_i) \varepsilon_i | V_i] = 0$ . Since data from subjects selected only to the first-stage sample contribute to the estimating equations (10) only through the function  $\phi_1(V_i, Z_i)$ , the efficient estimator of  $\alpha_0$  would be based solely on data from subjects selected to the second and/or third-stage samples and it would therefore disregard the recorded values of  $V_i$  from subjects in the first-stage sample only. We conclude that when data on  $Z_i$  are not collected then, asymptotically, the recorded data  $V_i$  from subjects not selected into the second or third-stage sample do not provide information about  $\alpha_0$ .

The functions  $h_{\text{eff}}$  and  $\phi_{\text{eff}}$  cannot, in general, be used for estimating  $\alpha_0$  since: (a) except in the special cases discussed in Section 6,  $h_{\text{eff}}$  does not exist in closed form in the sense that it cannot be explicitly represented as a function of the true distribution of the data, and (b)  $h_{\text{eff}}$  and  $\phi_{\text{eff}}$  depend on the unknown distribution of the data. In Section 6 we discuss a nearly efficient adaptive procedure that overcomes the difficulties described in (a) and (b).

### 5. $X_i$ and/or $Y_i$ missing by happenstance

In many epidemiological studies the data are missing by happenstance rather than by design, and thus the non-response probabilities  $\pi_{1i}$  and  $\pi_{2i}$  are unknown. For example, in the asthma study described in the introduction, respiratory health records may simply not exist on a subset of subjects, or some subjects may simply refuse to take a forced respiratory exam. In this setting, suppose that we continue to assume that the restrictions of the ‘three-stage’ model hold except for condition (6) that the selection probabilities are known. Instead, suppose that we correctly specify parametric models for the selection probabilities

$$\pi_{1i} = \pi_{1i}(\psi_0) \quad (12)$$

and

$$\pi_{2i} = \pi_{2i}(\psi_0) \quad (13)$$

where  $\pi_{1i}(\psi) \equiv \pi_1(V_i, Z_i; \psi)$  and  $\pi_{2i}(\psi) \equiv \pi_2(Y_i, V_i, Z_i; \psi)$  are smooth functions of a  $(r \times 1)$ -dimensional parameter  $\psi$  and of  $(V_i, Z_i)$  and  $(Y_i, V_i, Z_i)$  respectively. We can obtain consistent estimators of  $\alpha_0$  as follows. We first obtain  $\hat{\psi}$ , an efficient estimator of  $\psi_0$  in model (12)–(13). For example,  $\hat{\psi}$  is the consistent root of  $\sum_{i=1}^n S_{\psi,i}(\psi) = 0$  where  $S_{\psi,i}(\psi)$  is the contribution from the  $i$ th subject to the score for  $\psi$  in model (12)–(13). It is straightforward to show that

$$S_{\psi,i}(\psi) = \frac{\Delta_{1i} - \pi_{1i}(\psi)}{\pi_{1i}(\psi)} \phi_{\psi 1}(V_i, Z_i; \psi) + \frac{\Delta_{2i} - \pi_{2i}(\psi)}{\pi_{2i}(\psi)} \phi_{\psi 2}(Y_i, V_i, Z_i; \psi), \quad (14)$$

where

$$\phi_{\psi_1}(V_i, Z_i; \psi) = \pi_{1i} \frac{\partial \logit \pi_{1i}(\psi)}{\partial \psi} \quad (15)$$

and

$$\phi_{\psi_2}(Y_i, V_i, Z_i; \psi) = \pi_{2i} \frac{\partial \logit \pi_{2i}(\psi)}{\partial \psi} \quad (16)$$

Our estimators  $\hat{\alpha}(h, \phi; \hat{\psi})$  of  $\alpha_0$  solve

$$\sum_{i=1}^n D_i(x; h, \phi_1, \phi_2, \hat{\psi}) = 0, \quad (17)$$

where  $D_i(x; h, \phi_1, \phi_2, \psi)$  is defined like  $D_i(x; h, \phi_1, \phi_2)$  but with  $\pi_{1i}(\psi)$  and  $\pi_{2i}(\psi)$  instead of  $\pi_{1i}$  and  $\pi_{2i}$ . The following theorem states the asymptotic properties of  $\hat{\alpha}(h, \phi; \hat{\psi})$ .

**Theorem 4.** Let  $H(\gamma)^T = ([h(X, V)\varepsilon(x)]^T, S_\psi(\psi)^T)$ ,  $\gamma^T = (\alpha^T, \psi^T)$ , and  $\gamma = \alpha \times \psi$ , where  $\alpha$  and  $\psi$  are the parameter spaces of  $\alpha$  and  $\psi$ . Suppose that the regularity conditions 1, 2 and 4–9 of Theorem 1 hold with the new definition of  $H(\gamma)$ . Furthermore suppose that it also holds that  $\pi_1(V_i, Z_i; \psi) > \sigma > 0$  and  $\pi_2(Y_i, V_i, Z_i; \psi) > \sigma > 0$  almost surely for some  $\sigma$  and for all  $\psi \in \psi$ . If the assumptions of the ‘three-stage’ model hold and if (12) and (13) are correctly specified then with probability going to 1 Eq. (17) has a unique solution  $\hat{\alpha}(h, \phi; \hat{\psi})$ . This solution satisfies

- $\hat{\alpha}(h, \phi; \hat{\psi})$  is consistent;
- $\sqrt{n}\{\hat{\alpha}(h, \phi; \hat{\psi}) - \alpha_0\}$  has an asymptotic normal distribution with mean zero and variance  $I(h)^{-1} \text{Resid}\{D_i(\alpha_0; h, \phi), S_{\psi_i}(\psi_0)\} I(h)^{-1 \cdot T}$  that can be consistently estimated with  $\hat{I}(h)^{-1} \widehat{\text{Resid}}\{D_i(\alpha_0; h, \phi), S_{\psi_i}(\psi_0)\} \hat{I}(h)^{-1 \cdot T}$  where  $I(h)$  was defined in Theorem 1,  $\hat{I}(h) = n^{-1} \sum_i \partial D_i(x; h, \phi, \hat{\psi}) / \partial \alpha^T$ , for any random vectors  $A_i$  and  $B_i$   $\text{Resid}(A_i, B_i)$  is the residual variance from the population linear regression of  $A_i$  on  $B_i$ , i.e.  $\text{Resid}(A_i, B_i) = \text{var}(A_i) - \text{cov}(A_i, B_i) \text{var}(B_i)^{-1} \text{cov}(A_i, B_i)^T$  and  $\widehat{\text{Resid}}(A_i, B_i) = n^{-1} \sum_i (A_i - \hat{\beta} B_i)(A_i - \hat{\beta} B_i)^T$  with  $\hat{\beta}$  defined as the least squares coefficient in the linear regression of  $A_i$  on  $B_i$ ;
- $\hat{\alpha}(h, \phi; \hat{\psi})$  and  $\hat{\alpha}(h, \phi^*)$  are asymptotically equivalent, where  $\phi^* = \phi + E[D(h, \phi) S_\psi^T] \{\text{var}(S_\psi)\}^{-1} \phi_\psi$ , with  $S_\psi \equiv S_\psi(\psi_0)$ ,  $\phi_\psi \equiv (\phi_{\psi_1}(V, Z; \psi_0), \phi_{\psi_2}(Y, V, Z; \psi_0))$  and  $\phi_{\psi_1}(V, Z; \psi_0)$  and  $\phi_{\psi_2}(Y, V, Z; \psi_0)$  defined in (15) and (16);
- $\text{AVar}\{n^{1/2}[\hat{\alpha}(h, \phi) - \alpha_0]\} \geq \text{AVar}\{n^{1/2}[\hat{\alpha}(h, \phi, \hat{\psi}) - \alpha_0]\} \geq \text{AVar}\{n^{1/2}[\hat{\alpha}(h, \phi^h) - \alpha_0]\}$ , where  $\text{AVar}\{\delta_n\}$  denotes the variance of the asymptotic law of  $\delta_n$ ;
- given  $J$  nested correctly specified models,  $j = 1, \dots, J$ , for the missingness process ordered by the increasing dimension of the parameter vector  $\psi^{(j)}$ , the asymptotic variance of  $n^{1/2}\{\hat{\alpha}(h, \phi; \hat{\psi}^{(j)}) - \alpha_0\}$  is non-increasing in  $j$ ;
- if  $Y_i, V_i$  and  $Z_i$  are discrete and models (15) and (16) are saturated, then  $\hat{\alpha}(h, \phi; \hat{\psi})$  is asymptotically equivalent to  $\hat{\alpha}(h, \phi^h)$ .

Theorem 4 part (c) says that  $\hat{\alpha}(h, \phi; \hat{\psi})$  has, for some specific  $\phi^*$ , the same asymptotic distribution as  $\hat{\alpha}(h, \phi^*)$ , the solution to (10) that uses  $\phi^*$  and  $\pi$  known. Part (d) says that estimating the missingness probabilities efficiently under some correctly specified model for non-response when in fact they are known can never decrease the efficiency with which  $\alpha_0$  is estimated. This phenomenon has been noted by Robins et al. (1994) who also give a heuristic explanation for it. Part (e) says that we can improve the efficiency with which we estimate  $\alpha_0$  by adding more parameters to the models for the missingness probabilities. There exists, however, a lower bound for the asymptotic variance of  $\hat{\alpha}(h, \phi; \hat{\psi})$ . This bound, by Part (d), is equal to the asymptotic variance of  $\hat{\alpha}(h, \phi^h)$ . Part (f) says that when  $Y_i$ ,  $V_i$  and  $Z_i$  are discrete this lower bound is achieved by the asymptotic variance of the solution to (17) that uses the empirical estimates of the selection probabilities.

**Remark.** In Eq. (17) we could have used a  $n^{1/2}$ -consistent and asymptotically unbiased but inefficient estimator  $\hat{\psi}$  of  $\psi_0$  in model (12)–(13) (an estimator  $\hat{\psi}$  of  $\psi_0$  in model (12)–(13) is efficient if  $n^{1/2}(\hat{\psi} - \psi_0) = E[S_{\psi_i}(\psi_0)S_{\psi_i}(\psi_0)^T]^{-1}n^{-1/2}\sum_{i=1}^n S_{\psi_i}(\psi_0) + o_p(1)$ , and it is inefficient if this inequality is false). It can be shown that Eq. (17) that uses an inefficient  $\hat{\psi}$  will have a root  $\hat{\alpha}(h, \phi; \hat{\psi})$  that is consistent and asymptotically normal for estimating  $\alpha_0$ . However, the asymptotic variance of  $n^{1/2}(\hat{\alpha}(h, \phi; \hat{\psi}) - \alpha_0)$  will no longer be given by the formula in part (b) of Theorem 4. Furthermore, parts (c) and (d) are no longer true if  $\hat{\psi}$  is an inefficient estimator of  $\psi_0$  in model (12)–(13) and part (e) is false when  $\hat{\psi}^{(j)}$  are inefficient estimators of the parameter vectors  $\psi^{(j)}$ ,  $j = 1, \dots, J$ .

## 6. Adaptive estimation

The optimal functions  $\phi^h$  and  $h_{\text{eff}}$  are not useful for data analysis since they depend on the unknown distribution of the data. In this section we derive estimators  $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$  that use functions  $\hat{h}_{\text{eff}}$  and  $\hat{\phi}_{\text{eff}}$  calculated from the data and discuss their asymptotic properties. Our plan is as follows. We first consider the estimation of  $\phi^h$  for a fixed choice of  $h$ . Then we show how to obtain an iterative solution to the integral equation for  $h_{\text{eff}}$  in (11) under a fixed law of the data satisfying the restrictions of the model, and we discuss special cases in which an explicit solution of (11) can be obtained. Finally, we present an adaptive estimation procedure for locally semiparametric efficient estimation of  $\alpha_0$  under a fixed law. A locally semiparametric efficient estimator under model  $A$  at an additional restriction  $R$  is an estimator that attains the semiparametric variance bound when both  $A$  and  $R$  are true and remains consistent when  $A$  is true but  $R$  is false.

### 6.1. Adaptive estimation of $\phi^h$

For a fixed value of  $h$ , we now consider the estimation of  $\phi^h$ . Suppose we have specified the (possibly nonlinear) regression models

$$E\{h(X_i, V_i) \varepsilon_i(\alpha_0) | V_i, Z_i\} = e_1(V_i, Z_i; \lambda_{1,0}) \quad (18)$$

$$E\{h(X_i, V_i) \varepsilon_i(\alpha_0) | Y_i, V_i, Z_i\} = e_2(Y_i, V_i, Z_i; \lambda_{2,0}), \tag{19}$$

where  $e_1$  and  $e_2$  are known functions, smooth with respect to  $\lambda_1$  and  $\lambda_2$ , respectively. We estimate  $\lambda_{2,0}$  with  $\hat{\lambda}_2$  the (possibly non-linear) least squares estimator in the regression  $h(X_i, V_i) \varepsilon_i(\hat{\alpha})$  on  $(Y_i^T, V_i^T, Z_i^T)$  among subjects in the validation study, where  $\hat{\alpha} = \hat{\alpha}(h, 0)$  is the preliminary estimator of  $\alpha_0$  that uses  $\phi \equiv 0$ . Unfortunately, to estimate  $\lambda_{1,0}$  we cannot simply regress  $h(X_i, V_i) \varepsilon_i(\hat{\alpha})$  on  $(V_i^T, Z_i^T)$  using data on subjects in the validation sample since condition (4) does not imply that  $E\{h(X, V) \varepsilon | V, Z, \Delta_2 = 1\} = E\{h(X, V) \varepsilon | V, Z\}$ . However, it is straightforward to show that the inverse-probability-of-selection-weighted estimating equations

$$\sum_i \frac{\Delta_{2i}}{\pi_{1i}\pi_{2i}} u(V_i, Z_i) \{h(X_i, V_i) \varepsilon_i - e_1(V_i, Z_i; \lambda_1)\} = 0, \tag{20}$$

where  $u(V_i, Z_i)$  is any smooth vector function of  $(V_i^T, Z_i^T)$  of the same dimension as  $\lambda_1$ , have a solution  $\hat{\lambda}_1$  that is a consistent and asymptotically normal estimator of  $\lambda_{1,0}$ . Thus, we estimate  $\lambda_{1,0}$  with the solution  $\hat{\lambda}_1$  to (20) with  $\varepsilon_i$  replaced by  $\varepsilon_i(\hat{\alpha})$  where  $\hat{\alpha} \equiv \hat{\alpha}(h, 0)$ .

Our estimators of  $\phi_1^h$  and  $\phi_2^h$  are given by  $\hat{\phi}_1^h = e_1(V_i, Z_i; \hat{\lambda}_1)$  and  $\hat{\phi}_2^h = e_2(Y_i, V_i, Z_i; \hat{\lambda}_2)$ . It is straightforward to show that the estimator  $\hat{\alpha}(h, \hat{\phi}^h)$  has the same asymptotic distribution as  $\hat{\alpha}(h, \phi^\dagger)$  where  $\phi^\dagger = (\phi_1^\dagger, \phi_2^\dagger)$  is the probability limit of  $\hat{\phi}^h = (\hat{\phi}_1^h, \hat{\phi}_2^h)$  (see for example Robins et al., 1994, proof of Proposition 2.4). A consistent estimator of the asymptotic variance of  $\hat{\alpha}(h, \hat{\phi}^h)$  is provided by the variance estimator given in Theorem 1 but with  $\phi^h$  replaced by  $\hat{\phi}^h$ . Thus, when (18) and (19) are correctly specified,  $\hat{\alpha}(h, \hat{\phi}^h)$  is asymptotically equivalent to  $\hat{\alpha}(h, \phi^h)$ . Furthermore, as noted by Robins et al. (1994),  $\hat{\alpha}(h, \hat{\phi}^h)$  will be consistent for  $\alpha_0$  even if (18) or (19) are misspecified or incompatible with the restriction (1).

When  $Y_i, V_i$  and  $Z_i$  are discrete, the estimators

$$\hat{\phi}_2^h(y, v, z) = \left\{ \sum_i \Delta_{2i} I[(Y_i^T, V_i^T, Z_i^T) = (y^T, v^T, z^T)] \right\}^{-1} \times \left\{ \sum_i \Delta_{2i} I[(Y_i^T, V_i^T, Z_i^T) = (y^T, v^T, z^T)] h(X_i, V_i) \varepsilon_i(\hat{\alpha}) \right\} \tag{21}$$

and

$$\hat{\phi}_1^h(v, z) = \left\{ \sum_i \frac{\Delta_{2i}}{\pi_{1i}\pi_{2i}} I[(V_i^T, Z_i^T) = (v^T, z^T)] \right\}^{-1} \times \left\{ \sum_i \frac{\Delta_{2i}}{\pi_{1i}\pi_{2i}} I[(V_i^T, Z_i^T) = (v^T, z^T)] h(X_i, V_i) \varepsilon_i(\hat{\alpha}) \right\}, \tag{22}$$

where  $\hat{\alpha} = \hat{\alpha}(h, 0)$  and  $I(E)$  is the indicator of the event  $E$ , are guaranteed to converge in probability to  $\phi_2^h$  and  $\phi_1^h$  respectively. Thus, when  $Y_i, V_i$  and  $Z_i$  are discrete,  $\hat{\alpha}(h, \hat{\phi}^h)$  with  $\hat{\phi}^h = (\hat{\phi}_1^h, \hat{\phi}_2^h)$  defined in (21) and (22) is efficient for the specific choice of  $h$ .

6.2. Solution to the integral equation (11) by successive approximations

For any distribution  $f^*(Y, X, V, Z)$  allowed by the model satisfying restriction (1) we can obtain the solution  $h_{\text{eff}}$  to the integral equation (11) iteratively by the method of successive approximations as follows: (a) arbitrarily pick an initial value  $h_0(X, V)$ ; (b) calculate  $h_{m+1}(X, V)$  by evaluating the right-hand side of (11) at  $h_m(X, V)$  and taking expectations with respect to  $f^*(Y, X, V, Z)$ ; and (c) iterate until convergence. Using the same arguments as in Robins et al. (1994), it can be shown that  $h_m(X, V)$  converges to  $h_{\text{eff}}(X, V)$  as  $m \rightarrow \infty$ .

6.3. Closed-form solutions for  $h_{\text{eff}}$

We now consider some special cases in which there exists a closed-form expression for  $h_{\text{eff}}$ .

6.3.1. ‘Two-stage’ model when data on  $Z$  are not available

When data on  $Z$  are not available  $\phi_{1,\text{eff}} = 0$ . Thus, since in the ‘two-stage’ model the third term in (11) is zero because  $\pi_{2i} = 1$ , when data on  $Z$  are not available  $h_{\text{eff}}(X_i, V_i) = \{\partial g(X_i, V_i; \alpha_0) / \partial \alpha\} t(X_i, V_i)$ . Note that in this case, the semiparametric efficient estimator is just the complete-case estimator solving  $\sum_i \Delta_{1i} \{\partial g(X_i, V_i; \alpha_0) / \partial \alpha\} \{\text{var}(\varepsilon_i | X_i, V_i)\}^{-1} \varepsilon_i(\alpha) = 0$ .

6.3.2. Normal model

Suppose that data on  $Z$  are not available,  $Y$  is a scalar random variable, and consider the ‘three-stage’ model satisfying

$$g(X_i, V_i; \alpha_0) = \alpha_{01} + \alpha_{02} X_i, \tag{23}$$

$$\pi_{1i} \text{ is a constant } \rho_1, \pi_{2i} \text{ is a constant } \rho_2. \tag{24}$$

Define  $V^* = (1, V^T)^T$ . Suppose that in truth,

$$V \sim N(\mu, \Sigma), X|V \sim N(\gamma^T V^*, \Omega), \varepsilon|X, V \sim N(0, \sigma^2). \tag{25}$$

for some  $\mu = \mu_0, \gamma = \gamma_0, \Sigma = \Sigma_0, \Omega = \Omega_0$  and  $\sigma^2 = \sigma_0^2$ .

Let  $\mathcal{L}^F(\alpha, \eta; L)$  be the likelihood for the full data  $L = (Y^T, X^T, V^T)^T$  in the model satisfying (23) and (25) where  $\eta = (\mu, \Sigma, \gamma, \Omega, \sigma^2)$ . The observed-data likelihood in the fully parametric normal model satisfying restrictions (23)–(25) is  $\mathcal{L}(\alpha, \eta; \Delta_1, \Delta_2, L_{\text{obs}}) = \mathcal{L}^F(\alpha, \eta; L)^{\Delta_1 \Delta_2} \{\int \mathcal{L}^F dX\}^{\Delta_1(1-\Delta_2)} \{\int \int \mathcal{L}^F dX dY\}^{1-\Delta_1} \rho_1^{\Delta_1} (1-\rho_1)^{1-\Delta_1} \rho_2^{\Delta_1 \Delta_2} (1-\rho_2)^{\Delta_1(1-\Delta_2)}$ . The  $i$ th-subject’s contribution to the score equations for  $\theta = (\alpha^T, \eta^T)^T$  is then given by

$$\begin{aligned} S_{\theta,i}^{(NOR)} \equiv & \Delta_{1i} \Delta_{2i} \frac{\partial \log \mathcal{L}_i^F(\theta)}{\partial \theta} + \Delta_{1i} (1 - \Delta_{2i}) E_{\theta} \left\{ \frac{\partial \log \mathcal{L}_i^F(\theta)}{\partial \theta} \middle| Y_i, V_i \right\} \\ & + (1 - \Delta_{1i}) E_{\theta} \left\{ \frac{\partial \log \mathcal{L}_i^F(\theta)}{\partial \theta} \middle| V_i \right\}, \end{aligned} \tag{26}$$

where  $E_\theta$  denotes the expectation taken under the parameter value  $\theta$ . The solution  $\hat{\theta}_{MLE} = (\hat{\alpha}_{MLE}^T, \hat{\eta}_{MLE}^T)^T$  to  $\sum_i S_{\theta,i}^{(NOR)} = 0$  is the maximum likelihood estimator of  $\theta_0$ . In the Appendix we show that  $\hat{\alpha}_{MLE}$  remains consistent for estimating  $\alpha_0$  even when (23) and (24) hold but (25) is false. Thus,  $\hat{\alpha}_{MLE}$  is locally semiparametric efficient in the model ‘three-stage’ satisfying (23) and (24) at the additional restriction (25) with efficient score  $S_{\alpha,eff} = S_\alpha^{(NOR)} - E[S_\alpha^{(NOR)} S_\eta^{(NOR)T}] \{E[S_\eta^{(NOR)} S_\eta^{(NOR)T}]\}^{-1} S_\eta^{(NOR)}$ . Now, the closed-form expression  $h_{eff}(X, V) = E[S_{\alpha,eff} | Y = 1, X, V] - E[S_{\alpha,eff} | Y = 0, X, V]$  follows from the following lemma proved in the Appendix.

**Lemma 2.** *If  $\tilde{\alpha}$  is a regular asymptotically linear estimator of  $\alpha_0$  in the ‘three-stage’ model with influence function  $B = \Delta_1 \Delta_2 b_1(Y, X, V, Z) + \Delta_1(1 - \Delta_2) b_2(Y, V, Z) + (1 - \Delta_1)(1 - \Delta_2) b_3(V, Z)$ , then  $\tilde{\alpha}$  is asymptotically equivalent to  $\hat{\alpha}(h, \phi)$  (i.e.,  $n^{1/2}\{\tilde{\alpha} - \hat{\alpha}(h, \phi)\} = o_p(1)$ ), where  $\phi_1(V, Z) = b_3(V, Z)$ ,  $\phi_2(Y, V, Z) = \pi_1 b_2(Y, V, Z) + (1 - \pi_1) b_3(V, Z)$  and  $h = (h_1, \dots, h_i)$  with  $h_j(V, X) = E[B | Y = e_j, X, V] - E[B | Y = 0, X, V]$  and  $e_j$  and  $\mathbf{0}$  are vectors of the same dimension as  $Y$ ,  $e_j$  has  $j$ th element equal to 1 and all other elements equal to 0, and  $\mathbf{0}$  has all its elements equal to zero.*

6.3.3. *(Y, Z) Discrete in the ‘three-stage’ model*

We now show that when  $Y$  and  $Z$  are discrete,  $h_{eff}$  exists in closed form. First notice that in the integral equation (11),  $\phi_{1,eff} = E\{\phi_{2,eff} | V, Z\}$ , thus if we find a closed-form expression for  $\phi_{2,eff}$ , (11) provides a closed-form expression for  $h_{eff}$ . Upon right multiplying both members in (11) by  $\varepsilon$  and then taking conditional expectations with respect to  $(Y, V, Z)$  we obtain that  $\phi_{2,eff}$  solves

$$\begin{aligned} \phi_2(Y, V, Z) = & m(Y, V, Z) + E\{E[(1 - \pi_1)\pi_1^{-1} E(\phi_2 | V, Z) \varepsilon^T | X, V] t(X, V) \varepsilon | Y, V, Z\} \\ & + E\{E[(1 - \pi_2)(\pi_1 \pi_2)^{-1} \phi_2 \varepsilon^T | X, V] t(X, V) \varepsilon | Y, V, Z\}, \end{aligned} \tag{27}$$

where  $m(Y, V, Z) = E\{[\partial g(X, V; \alpha_0) / \partial \alpha] t(X, V) \varepsilon | Y, V, Z\}$ , and  $t(X, V)$  is defined in Theorem 3. In the Appendix we show that (27) has a unique solution.

Suppose for simplicity of notation that  $\alpha_0$  is an unknown scalar (for a  $q \times 1$  vector  $\alpha_0$  the following expression is obtained analogously for each of the  $q$  components of the vector  $\phi_2(Y, V, Z)$ ).

Suppose that  $Y$  takes  $J_1$  values  $y_1, \dots, y_{J_1}$  and  $Z$  takes  $J_2$  values  $z_1, \dots, z_{J_2}$ . Let  $J = J_1 \times J_2$  and with a slight abuse of notation define  $\phi_2(V)$  and  $m(V)$  to be the  $J \times 1$  vectors with  $j$ th element  $\phi_2(y_{j_1}, V, z_{j_2+1})$  and  $m(y_{j_1}, V, z_{j_2+1})$  respectively where  $j = j_2 J_1 + j_1$ ,  $0 \leq j_2 \leq J_2 - 1$  and  $1 \leq j_1 \leq J_1$ . In the Appendix we show that a closed-form expression for  $\phi_{2,eff}$  is given by

$$\phi_{2,eff}(V) = \{I_{J \times J} - q(V)\}^{-1} m(V), \tag{28}$$

where  $I_{J \times J}$  is the  $J \times J$  identity matrix and  $q(V)$  is a  $J \times J$  matrix, with  $(k, s)$  element

$$q_{ks}(V) = E \left\{ E \left[ \left( \frac{1 - \pi_1}{\pi_1} \right) \varepsilon^T \Pr(Y = y_{s_1} | V, Z) I(Z = z_{s_2+1}) P(Z | Y, X, V) | X, V \right] \right. \\ \left. \times t(X, V) [y_{k_1} - g(X, V)] | Y = y_{k_1}, V, Z = z_{k_2+1} \right\} \\ + E \left\{ \left( \frac{1 - \pi_2(y_{s_1}, V, z_{s_2+1})}{\pi_1(V, z_{s_2+1}) \pi_2(y_{s_1}, V, z_{s_2+1})} \right) [y_{s_1} - g(X, V)]^T \right. \\ \left. \times \Pr(Y = y_{s_1}, Z = z_{s_2+1} | X, V) t(X, V) [y_{k_1} - g(X, V)] \right. \\ \left. | Y = y_{k_1}, V, Z = z_{k_2+1} \right\},$$

where  $g(X, V) \equiv g(X, V; \alpha_0)$ ,  $k = k_2 J_1 + k_1$ ;  $s = s_2 J_1 + s_1$ ;  $1 \leq k_1, s_1 \leq J_1$ ; and  $0 \leq k_2, s_2 \leq J_2 - 1$ .

The function  $\phi_{1, \text{eff}}$  is then calculated as  $E\{\phi_{2, \text{eff}} | V, Z\}$ .

#### 6.3.4. $Z$ discrete in the 'two-stage' model

When restriction (5) holds the third term in the integral equation (11) vanishes. In this case if we find a closed-form expression for  $\phi_{1, \text{eff}}$ , then Eq. (11) provides a closed-form expression for  $h_{\text{eff}}$ . Upon right multiplying Eq. (11) by  $\varepsilon$  and then taking conditional expectations with respect to  $(V, Z)$  we obtain that  $\phi_{1, \text{eff}}$  solves

$$\phi_1(V, Z) = r(V, Z) + E\{E[(1 - \pi_1)\pi_1^{-1}\phi_1\varepsilon^T | X, V] t(X, V)\varepsilon | V, Z\}, \quad (29)$$

where  $r(V, Z) = E\{[\partial g(X, V; \alpha_0)/\partial \alpha] t(X, V)\varepsilon | V, Z\}$ , and  $t(X, V)$  is defined in Theorem 3 with  $\pi_2 = 1$ . In the Appendix we show that Eq. (29) has a unique solution. Using the notation of Section 6.3.3 and assuming, without loss of generality, that  $\alpha_0$  is a scalar, we obtain the closed-form expression for  $\phi_{1, \text{eff}} \equiv (\phi_{1, \text{eff}}(V, z_1), \dots, \phi_{1, \text{eff}}(V, z_{J_2}))^T$ ,

$$\phi_{1, \text{eff}}(V) = \{I_{J_2 \times J_2} - l(V)\}^{-1} r(V), \quad (30)$$

where  $r(V) = (r(V, z_1), \dots, r(V, z_{J_2}))^T$ , and  $l(V)$  is the  $J_2 \times J_2$  matrix with  $k, s$  entry  $l_{k,s}(V)$  equal to

$$l_{ks}(V) = E \left\{ E \left[ \left( \frac{1 - \pi_1}{\pi_1} \right) I(Z = z_s) P(Z | Y, X, V) \varepsilon^T | X, V \right] t(X, V) \varepsilon | V, Z = z_k \right\}.$$

#### 6.3.5. $X$ independent of $Y$ given $V$

Suppose that  $Y$  is conditionally independent of  $X$  given  $V$  and that auxiliary variables  $Z$  are not collected. Then, it is straightforward to show that for any  $h(X, V)$ ,  $\phi_1^h = 0$ , and  $\phi_2^h = E\{h(X, V) | V\} \varepsilon$ . Furthermore, since  $\varepsilon = Y - g(V; \alpha_0)$  does not

depend on  $X$ ,  $E\{(1 - \pi_2)(\pi_1 \pi_2)^{-1} \varepsilon \varepsilon^T | X, V\} = E\{(1 - \pi_2)(\pi_1 \pi_2)^{-1} \varepsilon \varepsilon^T | V\}$  and  $E\{\varepsilon \varepsilon^T (\pi_1 \pi_2)^{-1} | X, V\} = E\{\varepsilon \varepsilon^T (\pi_1 \pi_2)^{-1} | V\}$ . Thus, the integral equation (11) simplifies to

$$h_{\text{eff}} = \left\{ \frac{\partial g(X, V; \alpha_0)}{\partial \alpha} \right\} t(V) + E(h_{\text{eff}} | V) E \left\{ \frac{1 - \pi_2}{\pi_1 \pi_2} \varepsilon \varepsilon^T | V \right\} t(V) \tag{31}$$

where  $t(V) = E(\varepsilon \varepsilon^T / \pi_1 \pi_2 | V)^{-1}$ . Taking conditional expectations with respect to  $V$  in (31), solving for  $E(h_{\text{eff}} | V)$  and then replacing its solution in (31) we obtain

$$h_{\text{eff}} = \left\{ \frac{\partial g(X, V; \alpha_0)}{\partial \alpha} \right\} t(V) + E \left\{ \frac{\partial g(X, V; \alpha_0)}{\partial \alpha} \middle| V \right\} E \left[ \frac{\varepsilon \varepsilon^T}{\pi_1} \middle| V \right]^{-1} \times \left\{ 1 - E \left[ \frac{\varepsilon \varepsilon^T}{\pi_1} \middle| V \right] t(V) \right\},$$

which simplifies to

$$h_{\text{eff}} = h_{\text{eff}}^F \sigma^2(V) t(V) + E(h_{\text{eff}}^F | V) \{ \pi_1 - \sigma^2(V) t(V) \}, \tag{32}$$

where  $\sigma^2(V) = E(\varepsilon \varepsilon^T | V)$  and  $h_{\text{eff}}^F = \{ \partial g(X, V; \alpha_0) / \partial \alpha \} \sigma^2(V)^{-1}$  is the efficient  $h$  for estimating  $\alpha_0$  if no data are missing. When  $\pi_1 = 1$ , i.e.,  $Y_i$  is measured on all subjects, Eq. (32) agrees with the closed-form expression for  $h_{\text{eff}}$  found by Robins et al. (1994).

#### 6.4. Locally efficient adaptive estimation

We now describe an adaptive estimator of  $\alpha_0$  that is locally semiparametric efficient at a full-data parametric submodel satisfying the restriction (1) with likelihood  $\mathcal{L}^F(\alpha, \eta; Y, X, V, Z)$  indexed by  $\alpha$  and a finite dimensional parameter  $\eta$ . Our estimators are calculated as follows. Given a preliminary estimate  $\hat{\alpha} = \hat{\alpha}(h, \phi)$  for arbitrary choices of  $h$  and  $\phi$ ,

1. Estimate  $\eta$  by  $\hat{\eta}$  solving

$$\sum_i \left( \frac{\Delta_{2i}}{\pi_{1i} \pi_{2i}} \right) \frac{\partial \log \mathcal{L}^F}{\partial \eta}(\hat{\alpha}, \eta; Y_i, X_i, V_i, Z_i) = 0.$$

2. Solve the integral equation (11) by successive approximations as described in (6.2), (or explicitly if a closed-form solution exists), at the law  $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; Y, X, V, Z)$ . Denote the solution  $\hat{h}_{\text{eff}}$ . Evaluate  $\hat{\phi}_{\text{eff}}$  at the law  $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; Y, X, V, Z)$  to obtain  $\hat{\phi}_{\text{eff}}$ .

It is straightforward to show that under regularity conditions (see, for example, Robins et al. (1994, Proposition 2.4))  $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$  is asymptotically equivalent to  $\hat{\alpha}(h^\dagger, \phi^\dagger)$ , where  $h^\dagger$  and  $\phi^\dagger$  are the probability limits of  $\hat{h}_{\text{eff}}$  and  $\hat{\phi}_{\text{eff}}$ . Thus, if the true law of  $(Y, X, V, Z)$  is equal to  $\mathcal{L}^F(\alpha_0, \eta_0; Y, X, V, Z)$  for some  $\eta_0$ , then  $h^\dagger = h_{\text{eff}}$  and  $\phi^\dagger = \phi_{\text{eff}}$ . The estimator  $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$  is therefore locally semiparametric efficient at the parametric submodel  $\mathcal{L}^F(\alpha, \eta; Y, X, V, Z)$ .

## 7. Simulation studies

In order to examine the finite sample properties of our estimators, we conducted two simulation studies, one with  $Y_i$  a scalar binary variable and another with  $Y_i$  a scalar continuous variable. We considered a two-stage study in which auxiliary variables are observed at the first stage, the true variables of interest,  $Y_i$  and  $X_i$ , are simultaneously measured in the second-stage validation sample, that is  $\Delta_{1i} = \Delta_{2i} \equiv \Delta_i$  and  $\pi_{2i} = 1$ , and no covariates  $V_i$  are collected.

We generated, for each of 6000 subjects, a binary exposure  $X_i$  with  $P(X_i = 1) = 0.6$ ; an outcome  $Y_i$  with law  $f_1(Y_i|X_i)$  or  $f_2(Y_i|X_i)$  depending on the study.  $f_1(Y_i|X_i)$ , the probability density of the binary outcome of our first study, satisfied  $f_1(Y_i = 1|X_i) = \{1 + \exp(-\alpha_{01} - \alpha_{02}X_i)\}^{-1}$ , for our second study  $Y_i$  was a continuous variable with density  $f_2(Y_i|X_i) \sim \text{Normal}(\alpha_{01} + \alpha_{02}X_i; 1)$ , and in both studies  $\alpha_{01} = 0.07$  and  $\alpha_{02} = 0.5$ . We generated  $Z_i = (Z_i^X, Z_i^Y, Z_i^{XY}, Z_i^D)^T$ , a  $4 \times 1$  vector of binary variables with  $P(Z_i^X = 1|Y_i, X_i) = \{1 + \exp(0.73 - 3X_i)\}^{-1}$ ,  $P(Z_i^Y = 1|Y_i, X_i) = \{1 + \exp(0.73 - 3Y_i)\}^{-1}$ ,  $P(Z_i^{XY} = 1|Y_i, X_i) = \{1 + \exp(1.5 - 3X_i - 3Y_i)\}^{-1}$ , and  $P(Z_i^D = 1|Y_i, X_i) = \{1 + \exp(2 - 3X_i - 3Y_i)\}^{-1}$ . The auxiliaries  $Z_i^X$  and  $Z_i^Y$  were therefore conditionally independent of  $Y_i$  given  $X_i$  and of  $X_i$  given  $Y_i$  respectively and their conditional probability laws satisfied  $P(Z_i^X = 1|X_i = 1) = P(Z_i^Y = 1|Y_i = 1) = 0.91$  and  $P(Z_i^X = 0|X_i = 0) = P(Z_i^Y = 0|Y_i = 0) = 0.67$ . The selection indicators  $\Delta_i$  were generated according to  $P(\Delta_i = 1|Y_i, X_i, Z_i) = \{1 + \exp(2.25 - 3Z_i^D)\}^{-1}$  so that selection into the validation study depended only on the variable  $Z_i^D$ . The data were generated so that  $Z_i^X$  and  $Z_i^Y$  were independent predictors of  $X_i$  and  $Y_i$  respectively,  $Z_i^{XY}$  and  $Z_i^D$  were strong predictors of both  $X_i$  and  $Y_i$ , but  $Z_i^D$  was the only determinant of missingness. Each study was based on 1000 simulations.

Table 1 shows the results for study 1 with  $Y_i$  binary and Table 2 shows results for study 2 with  $Y_i$  continuous. Each table provides Monte Carlo averages and variances for the estimators considered, their asymptotic relative efficiencies compared to the semiparametric efficient estimator of row 9 calculated as the ratio of the Monte Carlo variances, and the estimated coverage probability of 95% confidence intervals. The estimated coverage probabilities from 1000 replications have a simulation accuracy of approximately 0.6%. For  $Y_i$  binary and for those estimators known to be consistent from the theory, we have also calculated their asymptotic variances and computed their ratio of the Monte Carlo variance to their asymptotic variance. These ratios are shown in the last column of Table 1.

The estimator  $\hat{\alpha}_{2, \text{FULL}}$  was calculated using all of the data generated on  $Y_i$  and  $X_i$ , i.e., irregardless of their value of the selection indicator  $\Delta_i$ .  $\hat{\alpha}_{2, \text{FULL}}$  solves  $\sum_{i=1}^n h(X_i) \varepsilon_i(\alpha) = 0$  with  $h(X_i) = (1, X_i)^T$  and it is the semiparametric efficient estimator of  $\alpha_{02}$  had  $Y_i$  and  $X_i$  been measured on all units in the sample (Chamberlain, 1987). The estimator  $\hat{\alpha}_{2, \text{NAIVE}}$  solves Eqs. (9) using  $h(X_i) = (1, X_i)^T$ . The estimators in rows 3–9 solve Eqs. (10) with  $h(X_i) = (1, X_i)^T$ . The estimator in row 3 uses  $\phi \equiv 0$ . The estimators in rows 4–9 each use  $\hat{\phi}^h$ , the estimate of  $\phi^h$  described in Section 6.1, from saturated models for  $E[h(X_i)\varepsilon_i|Z_i^*]$ , where  $Z_i^*$  is the subvector of  $Z_i$  containing the auxiliary

Table 1  
Simulation results for  $Y$  discrete. Sample size = 6000; number of samples = 1000. True  $\alpha_2 = 0.5$

Row	Estimator of $\alpha_2$	MC Ave	MC Var	ARE	Cover. Prob	MC Var/Asy Var
1	$\hat{\alpha}_{2, \text{FULL}}$	0.4995	0.00269	3.75	0.962	0.94
2	$\hat{\alpha}_{2, \text{NAIVE}}$	-0.4081	0.00927	1.09	0.0	
3	$\hat{\alpha}_2(h, 0)$	0.4988	0.01542	0.65	0.956	1.03
	<i><math>\hat{\alpha}_2(h, \hat{\phi}^h)</math> and <math>\hat{\phi}^h</math> calculated from the auxiliaries:</i>					
4	$Z^D$	0.5020	0.01474	0.69	0.960	1.03
5	$Z^D, Z^X$	0.5019	0.01350	0.75	0.954	1.01
6	$Z^D, Z^Y$	0.5023	0.01446	0.70	0.939	1.08
7	$Z^D, Z^{XY}$	0.5009	0.01077	0.94	0.946	1.05
8	$Z^D, Z^X, Z^Y$	0.5025	0.01295	0.78	0.941	1.06
9	$Z^D, Z^X, Z^Y, Z^{XY}$	0.4992	0.01010	1.00	0.933	1.10
	<i><math>\hat{\alpha}_2(h, 0; \hat{\psi})</math> and model <math>\pi(\psi)</math> includes:</i>					
10	$Z^{XY}$	-0.1345	0.00958	1.05	0.0	
11	$Z^D$	0.5003	0.01473	0.69	0.958	1.03
12	$Z^D, Z^X$	0.5002	0.01348	0.75	0.955	1.01
13	$Z^D, Z^Y$	0.5006	0.01446	0.70	0.939	1.08
14	$Z^D, Z^{XY}$	0.4988	0.01077	0.94	0.945	1.05
15	$Z^D, Z^X, Z^Y$	0.5007	0.01293	0.78	0.944	1.06
16	$Z^D, Z^X, Z^Y, Z^{XY}$ (main effects only)	0.4996	0.01404	0.72	0.953	
17	$Z^D, Z^X, Z^Y, Z^{XY}$	0.4974	0.01008	1.00	0.938	1.10

variables listed in each respective row. Since  $X_i$  is binary, then any function of  $X_i$  is linear in  $X_i$ . In particular,  $h_{\text{eff}}(X_i)^T = (h_{\text{eff}1}(X_i)^T, h_{\text{eff}2}(X_i)^T)$  where  $h_{\text{eff}j}(X_i) = c_{1j} + c_{2j}X_i$  for some constants  $c_{1j}$  and  $c_{2j}$ ,  $j = 1, 2$ . Then  $h_{\text{eff}}(X_i) = C h(X_i)$  for the matrix  $C$  with entries  $c_{kj}$ ,  $k, j = 1, 2$ . Thus, each of the estimators in rows 4–9 would be semiparametric efficient if only the specifically listed auxiliary variables were available. The estimators in rows 10–17 solve Eqs. (10) that use  $\phi = 0$  and the true  $\pi_i$  replaced by an estimate of it. The estimator in row 10 uses  $\pi_i$  estimated from the misspecified model that assumes that  $\Delta_i$  follows a logistic regression on  $Z_i^{XY}$ . The estimators in rows 11–15 and row 17 use  $\pi_i$  estimated from saturated logistic regression models of  $\Delta_i$  on the auxiliary variables listed in each row, i.e., each model contains linear terms of each auxiliary variable as well as all first- and higher-order products. The estimator in row 16 used  $\pi_i$  estimated from a model that included only linear terms of the listed variables.

In the 1000 replications, the validation sample sizes were 51 and 44% of the sample sizes of the first-stage study for  $Y_i$  binary and  $Y_i$  continuous, respectively. In both tables, the complete-case estimator of row 2 and the estimator of row 10 that estimated  $\pi_i$  from a misspecified model are severely biased. A heuristic explanation of the smaller bias of the estimator in row 10 compared to that in row 2 is that: (a) the estimating Eq. (9) used  $\pi_i = 1$  for all  $i$ , and (b)  $\hat{\pi}_i$  is a better approximation to the true

Table 2  
Simulation results for  $Y$  continuous. Sample size = 6000; number of samples = 1000.  
True  $\alpha_2 = 0.5$

Row	Estimator of $\alpha_2$	MC Ave	MC Var	ARE	Cover. Prob
1	$\hat{\alpha}_{2, \text{FULL}}$	0.4998	0.00069	4.39	0.953
2	$\hat{\alpha}_{2, \text{NAIVE}}$	0.1212	0.00181	1.67	0.0
3	$\hat{\alpha}_2(h, 0)$	0.4987	0.00416	0.73	0.944
<i><math>\hat{\alpha}_2(h, \hat{\phi}^h)</math> and <math>\hat{\phi}^h</math> calculated from the auxiliaries:</i>					
4	$Z^D$	0.4999	0.00416	0.73	0.944
5	$Z^D, Z^X$	0.5006	0.00357	0.85	0.949
6	$Z^D, Z^Y$	0.4993	0.00400	0.76	0.938
7	$Z^D, Z^{XY}$	0.5000	0.00371	0.82	0.952
8	$Z^D, Z^X, Z^Y$	0.4997	0.00328	0.92	0.947
9	$Z^D, Z^X, Z^Y, Z^{XY}$	0.4998	0.00303	1.00	0.955
<i><math>\hat{\alpha}_2(h, 0; \hat{\psi})</math> and model <math>\pi(\psi)</math> includes:</i>					
10	$Z^{XY}$	0.2653	0.00234	1.29	0.003
11	$Z^D$	0.4993	0.00416	0.73	0.946
12	$Z^D, Z^X$	0.5000	0.00358	0.85	0.950
13	$Z^D, Z^Y$	0.4988	0.00400	0.76	0.940
14	$Z^D, Z^{XY}$	0.4992	0.00371	0.82	0.953
15	$Z^D, Z^X, Z^Y$	0.4992	0.00329	0.92	0.949
16	$Z^D, ZX, Z^Y, Z^XY$ (main effects only)	0.4992	0.00395	0.77	0.945
17	$Z^D, Z^X, Z^Y, Z^{XY}$	0.4992	0.00304	1.00	0.954

$\pi_i$  than the constant 1, even when  $\hat{\pi}_i$  is calculated from the misspecified logistic regression model of  $\Delta_i$  on  $Z_i^{XY}$ . This is possibly due to the fact that  $Z_i^{XY}$  is a better proxy for  $Z_i^D$  than a constant-valued variable. All other estimators considered were unbiased.

The almost identical asymptotic variances of the estimators in rows 4–9 compared to those in rows 11–15 and 17 using the same auxiliary variables is as predicted by the theory since,  $\hat{\alpha}_2(h, \hat{\phi}^h)$  and  $\hat{\alpha}_2(h, 0; \hat{\psi})$  are asymptotically equivalent. To verify that  $\hat{\alpha}_2(h, \hat{\phi}^h)$  and  $\hat{\alpha}_2(h, 0; \hat{\psi})$  are asymptotically equivalent first note that Theorem 4 part (f) implies that in a ‘two-stage’ model, i.e., when  $\pi_{2i} = 1$ ,  $\hat{\alpha}_2(h, \hat{\phi}^h)$  and  $\hat{\alpha}_2(h, 0; \hat{\psi})$  are asymptotically equivalent for any variable  $Y$ , not just for  $Y$  discrete. The assertion now follows since, as noted in Section 6.1,  $\hat{\alpha}(h, \hat{\phi}^h)$  and  $\hat{\alpha}(h, \phi^h)$  are asymptotically equivalent because  $\hat{\phi}^h$  is estimated from a saturated model for the conditional mean of  $h(X_i) \varepsilon_i$  given the auxiliaries. Since  $h_{\text{eff}}(X_i) = C h(X_i)$  for some constant  $2 \times 2$  matrix  $C$ , the estimators in rows 4–9 are semiparametric efficient for estimating  $\alpha_0$  when the only auxiliary variables available are those listed in each respective row. Thus, the differences among their Monte Carlo variances are exclusively due to the different amounts of information about  $\alpha_{02}$  carried by the available auxiliary variables in each row. For

example, a comparison of the asymptotic variances in rows 7 and 8 indicates that  $Z_i^{XY}$  carries more information about  $\alpha_{02}$  than the two independent predictors  $Z_i^X$  and  $Z_i^Y$  of  $X_i$  and  $Y_i$  together. The almost 50% increase in efficiency of the estimator in row 9 compared to the estimator in row 4 illustrates the importance of collecting auxiliary variables that are predictors of the missing data  $X$  and  $Y$  even when they are not determinants of missingness.

A comparison of the asymptotic variances of the estimators in rows 3 and 11–17 illustrates the efficiency gains obtained by estimating  $\pi_i$  even when it is known. A comparison of rows 16 and 17 indicates the efficiency improvements in the estimators of  $\alpha_0$  obtained by augmenting a linear logistic model for  $\pi_i$  with first- and higher-order interactions of the auxiliary variables. The last column of Table 1 indicates that the asymptotic variances of the estimators of  $\alpha_0$  are well-approximated by their finite-sample variance.

## 8. Final remarks

In this paper we have described a class of consistent and asymptotically normal estimators of  $\alpha_0$  when both outcomes and a subset of the covariates are not always observed and auxiliary variables are measured in all study subjects. Our methods allow for the probability of selection into the validation sample to depend on the values of the auxiliary variables. Our approach requires selection probabilities that are bounded away from zero and that are either known or can be parametrically modeled. As opposed to the fully parametric approach, our methods provide consistent estimators of  $\alpha_0$  under any law of the covariates and any conditional law of the auxiliary variables given the outcomes and covariates.

In particular, our estimators will be consistent even if Eq. (2) describing the nature of the noise in  $Z_i$  does not hold. Eq. (2), however, provides additional restrictions on the joint law of the data that are informative about  $\alpha_0$ . That is, it is possible to show that the semiparametric variance bound for estimating  $\alpha_0$  in the model defined by the ‘three-stage’ model restrictions and Eq. (2) is smaller than the asymptotic variance of  $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$ . Thus, if Eq. (2) is known to be true, it is possible to construct estimators that are more efficient than  $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$ . These estimators, however, will be inconsistent for estimating  $\alpha_0$  if Eq. (2) fails to be true.

Finally, when the missingness probabilities are unknown, as an alternative to using  $\pi_i(\hat{\psi})$ , we could have replaced  $\pi_i(\hat{\psi})$  by a completely non-parametric kernel regression estimator to protect against misspecification bias. Although a detailed consideration of the large sample properties of the resulting estimator of  $\alpha_0$  is beyond the scope of this paper, under sufficient regularity conditions and with the bandwidth appropriately chosen, this estimator can be shown to be asymptotically normal with asymptotic variance equal to that of  $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$ .

## Appendix

In this appendix we outline the proofs of the theorems and lemmas in this article. Because these proofs use arguments that are, in many cases, identical to those used in the proofs of the results in Robins et al. (1994), whenever possible, we prove our results by referring to the specific proofs in Robins et al. (1994) and specifying the appropriate changes in notation so that their arguments apply to our results. We start with the proof of Theorem 4. Theorem 1 is a special case of Theorem 4 parts (a) and (b) and its proof is therefore omitted.

**Proof of Theorem 4.** With the new definition of  $S_\psi(\psi)$ , the proof of parts (a)–(e) is identical to the proof of Proposition 6.1 of Robins et al. (1994). Part (f) is a corollary to part (e) since when  $Y_i$ ,  $V_i$  and  $Z_i$  are discrete there is only a finite number of correctly specified models  $\pi_{1i}$  and  $\pi_{2i}$ , and by part (e) the lower bound for the asymptotic variance of  $\hat{\alpha}(h, \phi; \hat{\psi})$  is achieved when the models  $\pi_{1i}(\psi)$  and  $\pi_{2i}(\psi)$  are saturated in  $(Y_i, X_i, V_i)$ .  $\square$

**Proof of Theorem 2.** The proof is identical to the proof of Proposition 3.2 of Robins et al. (1994) by defining  $A_i(\phi) = (\Delta_{1i} - \pi_{1i})\pi_{1i}^{-1}\phi_1(V_i, Z_i) + (\Delta_{2i} - \pi_{2i}\Delta_{1i})(\pi_{1i}\pi_{2i})^{-1}\phi_2(Y_i, V_i, Z_i)$  and noting that Eq. (4) implies that the missing data patterns are monotone.  $\square$

**Proof of Lemma 1.** Let  $D(h, \phi) = D(\alpha_0; h, \phi)$ . It follows from Theorem 1 that the functions  $\phi_1^h$  and  $\phi_2^h$  minimizing the asymptotic variance of  $\hat{\alpha}(h, \phi)$  minimize the variance of  $D(h, \phi)$ . Now, write

$$D(\alpha; h, \phi) = D^F(\alpha; h) + J(\alpha; h, \phi),$$

with  $D^F(\alpha; h) = h(X, V)\varepsilon(\alpha)$  and  $J(\alpha; h, \phi) = (\Delta_1 - \pi_1\pi_1^{-1})[D^F(\alpha; h) - \phi_1(V, Z)] + (\Delta_2 - \pi_2\Delta_1)(\pi_1\pi_2)^{-1}[D^F(\alpha; h) - \phi_2(Y, V, Z)]$ . Further, by (3), (4) and (7), for all  $\alpha$ ,  $(\Delta_1 - \pi_1)\pi_1^{-1}[D^F(\alpha; h) - \phi_1(V, Z)]$  has mean 0 given  $(V, Z)$  and  $(\Delta_2 - \pi_2\Delta_1)(\pi_1\pi_2)^{-1}[D^F(\alpha; h) - \phi_2(Y, V, Z)]$  has mean 0 given  $(Y, V, Z)$  and thus they are both uncorrelated with  $D^F(\alpha; h)$ . Hence, with the definition  $A^{\otimes 2} = AA^T$  for any  $A$ , we have

$$\begin{aligned} \text{Var}[D(\alpha; h, \phi)] &= \text{Var}[D^F(\alpha; h)] \\ &\quad + E\{(1 - \pi_1)\pi_1^{-1}E\{[D^F(\alpha; h) - \phi_1]^{\otimes 2}|V, Z\}\} \\ &\quad + E\{(1 - \pi_2)(\pi_1\pi_2)^{-1}E\{[D^F(\alpha; h) - \phi_2]^{\otimes 2}|Y, V, Z\}\}, \end{aligned}$$

which is minimized in the positive definite sense at  $\phi_1 = E[D^F(\alpha; h)|V, Z]$  and  $\phi_2 = E[D^F(\alpha; h)|Y, V, Z]$ .  $\square$

**Proof of Theorem 3.** The conditional mean model (1) coincides with model ‘full’ of Robins et al. (1994). Furthermore, by (3) and (4), the missing data patterns are

monotone. Thus, Propositions 8.1(e) and 8.3(b) of Robins et al. (1994) imply that  $h_{\text{eff}}$  solves  $S_{\text{eff}}^F = \Pi[m^{-1}(D^*)|A^{F,\perp}]$  with  $D^* \equiv h(X)\varepsilon$  and  $\Pi[\cdot|\cdot]$ ,  $m^{-1}(\cdot)$  and  $A^{F,\perp}$  defined as in Robins et al. (1994). Hence, Propositions 8.2(e) and 8.3(a) of Robins et al. (1994) imply that  $S_{\text{eff}}^F$  solves

$$S_{\text{eff}}^F = E \left\{ \left[ \frac{h(X)\varepsilon}{\pi_1\pi_2} - \frac{1-\pi_1}{\pi_1} E[h(X)\varepsilon|V, Z] - \frac{1-\pi_2}{\pi_1\pi_2} E[h(X)\varepsilon|Y, V, Z] \right] \varepsilon | X \right\} \times \text{Var}(\varepsilon|X)^{-1} \varepsilon \tag{A.1}$$

Eq. (11) follows after post-multiplying both members of (A.1) by  $\varepsilon^T$  and then taking conditional expectations with respect to  $X$ .

**Proof that  $\hat{\alpha}_{\text{MLE}}$  is a regular estimator of  $\alpha_0$ .** Let  $\eta^\dagger = (\gamma^\dagger, \Omega^\dagger, \sigma^{2\dagger}, \mu^\dagger, \Sigma^\dagger)$  represent a set of nuisance parameters defined as follows:  $\gamma^\dagger = E[X, V^{*\top}] \{E[V^* V^{*\top}]\}^{-1}$ ,  $\Omega^\dagger = E[(X - \gamma^{\dagger\top} V)^{\otimes 2}]$ ,  $\sigma^{2\dagger} = E\{[Y - (\alpha_{0,0} + \alpha_{0,1} X)]^2\}$ ,  $\mu^\dagger = E(V)$ , and  $\Sigma^\dagger = E[(V - \mu^{\dagger\top})^{\otimes 2}]$ , where expectations are taken under any law  $f^*$  of the data satisfying the restrictions of the ‘three-stage’ model. By definition of  $\gamma^\dagger$ ,  $E\{(X - \gamma^{\dagger\top} V^* V^{*\top})\} = 0$ , and therefore  $E(X - \gamma^{\dagger\top} V^*) = 0$  because the first element of  $V^*$  is 1. Thus,

$$E(X) = \gamma^{\dagger\top} E(V^*). \tag{A.2}$$

Now, to show that  $\hat{\alpha}_{\text{MLE}}$  is a consistent estimator of  $\alpha_0$ , it is enough to show that  $E[S_{\theta,i}^{\text{NOR}}(\alpha_0, \eta^\dagger)] = 0$  with expectation taken with respect to  $f^*$ . To show this, simply write the score for  $\theta$  in a normal model evaluated at  $\eta^\dagger$  and check that by definition of  $\eta^\dagger$  and by (A.2) it has mean zero at the law  $f^*$ .  $\square$

**Proof of Lemma 2.** It follows from Propositions 8.1 (c2) and 8.2(c) of Robins et al. (1994) that there exist  $h$  and  $\phi$  such that

$$B_i = D_i(h, \phi). \tag{A.3}$$

Evaluating (A.3) at  $\Delta_{1i} = \Delta_{2i} = 0$  we get  $\phi_1(V, Z) = b_3(V, Z)$ . Evaluation of (A.3) at  $\Delta_{1i} = 1$  and  $\Delta_{2i} = 0$  gives  $\phi_2(Y, V, Z) = \pi_1 b_2(Y, V, Z) + (1 - \pi_1) b_3(V, Z)$ . The proof of Lemma 2 is concluded after noticing that  $E\{D(h, \phi)|Y = e_j, X, V\} - E\{D(h, \phi)|Y = 0, X, V\}$  is equal to the  $j$ th column of the  $p \times T$  matrix function  $h(X, V)$  and calculating the same difference of conditional expectations on the left-hand side of (A.3).  $\square$

**Proof of Eqs. (28) and (30) for discrete  $Y$  and  $Z$ .** Eq. (28) is equivalent to

$$\phi_{2,\text{eff}}(V) - q(V) \phi_{2,\text{eff}}(V) = m(V),$$

so it is enough to show with  $k = k_2 J_1 + k_1$ ,

$$\begin{aligned} & [q(V)\phi_{2,\text{eff}}(V)]_k \\ &= E\{E[(1 - \pi_1)\pi_1^{-1} E(\phi_2|V, Z)\varepsilon^T|X, V]t(X, V)\varepsilon|Y = y_{k_1}, V, Z = z_{k_2+1}\} \\ & \quad + E\{E[(1 - \pi_2)(\pi_1\pi_2)^{-1}\phi_2\varepsilon^T|X, V]t(X, V)\varepsilon|Y = y_{k_1}, V, Z = z_{k_2+1}\}, \end{aligned}$$

where  $[q(V)\phi_{2,\text{eff}}(V)]_k$  denotes the  $k$ th element of the vector  $q(V)\phi_{2,\text{eff}}(V)$ ,  $k = 1, \dots, J$ . But with  $q(V)_k$  denoting the  $k$ th row of the  $J \times J$  matrix  $q(V)$ , we have  $[q(V)\phi_{2,\text{eff}}(V)]_k = q(V)_k\phi_{2,\text{eff}}(V)$ .

The  $s$ th element of  $\phi_{2,\text{eff}}(V)$  is equal by definition to  $\phi_{2,\text{eff}}(y_{s_1}, V, z_{s_2+1})$  where  $s = s_1 J_1 + s_2$ ,  $0 \leq s_2 \leq J_2 - 1$  and  $1 \leq s_1 \leq J_1$  so,

$$\begin{aligned} & q(V)_k\phi_{2,\text{eff}}(V) \\ &= \sum_{s_1} \sum_{s_2} E \left\{ E \left( \frac{1 - \pi_1}{\pi_1} \right) \varepsilon^T \Pr(Y = y_{s_1}|V, Z) I(Z = z_{s_2+1}) P(Z|Y, X, V) \right. \\ & \quad \left. \phi_2(y_{s_1}, V, z_{s_2+1})|X, V \right\} t(X, V) [y_{k_1} - g(X, V)] | Y = y_{k_1}, V, Z = z_{k_2+1} \Big\} \\ & \quad + E \left\{ \left( \frac{1 - \pi_2(y_{s_1}, V, z_{s_2+1})}{\pi_1(V, z_{s_2+1})\pi_2(y_{s_1}, V, z_{s_2+1})} \right) [y_{s_1} - g(X, V)]^T \right. \\ & \quad \left. \Pr(Y = y_{s_1}, Z = z_{s_2+1}|X, V) \phi_2(y_{s_1}, V, z_{s_2+1}) t(X, V) [y_{k_1} - g(X, V)] \right. \\ & \quad \left. | Y = y_{k_1}, V, Z = z_{k_2+1} \right\}. \end{aligned}$$

But,

$$\begin{aligned} & E \left\{ \sum_{s_1} \sum_{s_2} (1 - \pi_1)\pi_1^{-1} \varepsilon^T \Pr(Y = y_{s_1}|V, Z) I(Z = z_{s_2+1}) P(Z|Y, X, V) \right. \\ & \quad \left. \times \phi_2(y_{s_1}, V, z_{s_2+1})|X, V \right\} \\ &= E \left\{ \sum_{s_2} [1 - \pi_1(V, z_{s_2+1})] \pi_1(V, z_{s_2+1})^{-1} \varepsilon^T E(\phi_2|V, Z = z_{s_2+1}) \right. \\ & \quad \left. \times P(Z = z_2|Y, X, V)|X, V \right\} \\ &= E\{E[(1 - \pi_1)\pi_1^{-1} \varepsilon^T E(\phi_2|V, Z)|Y, X, V]|X, V\} \\ &= E\{(1 - \pi_1)\pi_1^{-1} \varepsilon^T(\phi_2|V, Z)|X, V\}, \end{aligned}$$

and

$$\sum_{s_1} \sum_{s_2} \left( \frac{1 - \pi_2(y_{s_1}, V, z_{s_2+1})}{\pi_1(V, z_{s_2+1}) \pi_2(y_{s_1}, V, z_{s_2+1})} \right) [y_{s_1} - g(X, V)]^T P(Y = y_{s_1}, Z = z_{s_2+1} | X, V) \\ \times \phi_2(y_{s_1}, V, z_{s_2+1}) = E \left[ \left( \frac{1 - \pi_2}{\pi_1 \pi_2} \right) \varepsilon^T \phi_2 | X, V \right],$$

which proves the result.  $\square$

The proof of equation (30) follows by an analogous argument.

**Proof that Eqs. (27) and (29) have a unique solution.** The uniqueness of the solutions of Eqs. (27) and (29) follows from the uniqueness of the solution  $h_{\text{eff}}(X, V)$  of Eq. (11) using arguments identical to those used to prove the uniqueness of Eq. (27) in the paper of Robins et al. (1994). The uniqueness of  $h_{\text{eff}}(X, V)$ , on the other hand, follows by identical arguments as those used in Robins et al. (1994) to show the uniqueness of their  $h_{\text{eff}}(X^*)$ .

### Acknowledgements

This work was conducted as part of Christina Holcroft's doctoral dissertation.

### References

- Begun, J.M., Hall, W.J., Huang, W.M., Wellner, J.A., 1983. Information and Asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* 11, 432–452.
- Breslow, N.E., Cain, K.C., 1988. Logistic regression for two-stage case-control data. *Biometrika* 75, 11–20.
- Carroll, R.J., Wand, M.P., 1991. Semi-parametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B*, 53, 573–587.
- Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* 34, 305–324.
- Cosslett, S.R., 1981. Efficient estimation of discrete choice models. In: C.F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data With Econometric Applications*. MIT Press, Cambridge, MA, pp. 51–111.
- Flanders, W.D., Greenland, S., 1991. Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med.* 10, 739–747.
- Holcroft, C.A., Rotnitzky, A., Spiegelman, D., 1995. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified, Christina Holcroft's Doctoral Dissertation, Department of Biostatistics, Harvard School of Public Health.
- Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 663–685.
- Imbens, G.W., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* 60, 1187–1214.
- Kalbfleisch, J.D., Lawless, J.F., 1988. Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.* 7, 149–160.
- Manski, C.F., Lerman, S., 1977. The estimation of choice probabilities from choice-based samples. *Econometrica* 45, 1977–1988.

- Manski, C.F., McFadden, D., 1981. Alternative estimators and sample designs for discrete choice analysis. In: *Structural Analysis of Discrete Data With Econometric Applications*, Manski, C.F., McFadden, D. (Eds.) MIT Press, Cambridge, MA, pp. 2–50.
- Pepe, M.S., 1992. Inference using surrogate outcome data and a validation sample. *Biometrika* 79, 355–65.
- Pepe, M.S., Fleming, T.R., 1991. A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* 86, 108–113.
- Pepe, M.S., Reilly, M., Fleming, T.R., 1994. Auxiliary outcome data and the mean-score method. *J. Statist. Plann. Inference* 42, 137–160.
- Reilly, M., Pepe, M.S., 1995. A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika* 82, 299–314.
- Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 89, 846–866.
- Rotnitzky, A., Robins, J.M., 1995a. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 82, 805–820.
- Rotnitzky, A., Robins, J.M., 1995b. Efficient semiparametric estimation with missing outcomes and surrogate data. *Scand. J. Statist.*, under review.
- Rubin, D.B., 1976. Inference and missing data. *Biometrika* 63, 581–592.
- Tosteson, T.D., Ware, J.H., 1990. Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika* 77, 11–21.
- Zhao, L.P., Lipsitz, S., 1992. Design and analysis of two-stage studies. *Statist. Med.* 11, 769–782.