Biometrika (2015), **102**, 3, *pp.* 739–746
Printed in Great Britain

doi: 10.1093/biomet/asv025
Advance Access publication 25 June 2015

Semiparametric causal inference in matched cohort studies

BY E. H. KENNEDY

Department of Biostatistics and Epidemiology, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
kennedye@mail.med.upenn.edu

A. SJÖLANDER

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,
171 77 Stockholm, Sweden
arvid.sjolander@ki.se

AND D. S. SMALL

Department of Statistics, The Wharton School, University of Pennsylvania,
Philadelphia, Pennsylvania 19104, U.S.A.
dsmall@wharton.upenn.edu

SUMMARY

Odds ratios can be estimated in case-control studies using standard logistic regression, ignoring the outcome-dependent sampling. In this paper we discuss an analogous result for treatment effects on the treated in matched cohort studies. Specifically, in studies where a sample of treated subjects is observed along with a separate sample of possibly matched controls, we show that efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators, which ignore the matching and exposure-based sampling. This is not the case for general average effects. We also show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and derive the optimal number of matches for a given set of matching variables. We illustrate our results via simulation and in a matched cohort study of the effect of hysterectomy on the risk of cardiovascular disease.

Some key words: Biased sampling; Doubly robust; Effect on the treated; Efficient influence function; Study design.

1. INTRODUCTION

In this paper we consider matched cohort studies in which a sample of treated subjects is observed along with a separate sample of possibly matched controls. Such studies are particularly useful in settings where the treatment is relatively uncommon and it is expensive to collect either the outcome data or the full set of covariates. These designs are also widely used; according to PubMed the number of articles including both terms “matched” and “cohort” has increased every year since 2000, and totals 19 581 as of 8 January 2015. For example, Ingelsson et al. (2011) used a matched cohort design to estimate the effect of hysterectomy on the risk of cardiovascular disease. They first identified all Swedish women who underwent hysterectomies between 1973 and 2003 using the Swedish Inpatient Register, and then for each of these women matched three additional women who did not have a hysterectomy but who were the same age and lived in the same county. In this study it was difficult to collect outcome data about cardiovascular events, as well as additional covariate information such as socioeconomic status, because linkage to numerous additional national health registers was required. More examples and general discussion of matched cohort studies can be found in Jewell (2003) and Rothman et al. (2008).

Matched cohort studies are most often used for estimating treatment effects on the treated. These effects can be of more interest than average effects, especially when treatment is relatively rare and some subjects are very unlikely to receive it. A primary contribution of this paper is to show that effects on the treated can be estimated in matched cohort studies using standard methods, ignoring the study design; this is a cohort study analog of the famous odds ratio result for case-control studies (Anderson, 1972; Prentice & Pyke, 1979). To the best of our knowledge, this fact has never before been mentioned in the literature. It means that, for example, even though propensity scores are not identified in matched cohort designs, usual semiparametric, e.g., propensity score-based, doubly robust, estimators of the effect on the treated can be applied without modification, and without requiring external information about treatment prevalence or matched covariate distributions. Thus much of the important literature on semiparametric estimation of effects on the treated (Heckman et al., 1997; Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003; Imbens, 2004; Abadie & Imbens, 2006; Kline, 2011) is also relevant for matched cohort studies, even though this work has mostly focused on simple random sampling.

A number of authors have considered causal inference in matched cohort study settings, but none seem to have mentioned the above result. Heckman & Todd (2009) gave some justification for using the propensity score in exposure-stratified studies without matching, but did not discuss semiparametric theory or double robustness. Tchetgen Tchetgen & Rotnitzky (2011) developed semiparametric theory and doubly robust estimators for the conditional odds ratio but did not consider general marginal effects or efficiency across study designs. Sjölander et al. (2012) and Sjölander & Greenland (2013) discussed using likelihood-based regression methods, but did not consider using propensity scores. van der Laan et al. (2013) examined cohort studies for community-based interventions, but required external information beyond the sample.

2. Set-up

We consider the following study set-up. Covariates L and outcome Y are observed for n_1 treated subjects along with n_0 controls, where $n_0 = kn_1$ is fixed so that k controls are selected for each treated subject. In addition the controls can be matched to the treated on a subset of discrete covariates $W \subseteq L$. We use $\bar{W} = L \setminus W$ to denote the set of covariates not used in matching, so that $L = (W, \bar{W})$. The observed data are $(Z_1, \ldots, Z_n)$ with $Z = (L, A, Y)$ and A an indicator of treatment, where by design we have that $\sum_i A_i = n_1$, $\sum_i (1 - A_i) = n_0 = kn_1$, and $W_i = W_j$ if subjects i and j are matched. If there is no matching so that $W = \emptyset$ and $\bar{W} = L$ then we simply observe two separate random samples of treated and control subjects.

The main statistical issue in a matched cohort study is the fact that the observations are not independent and identically distributed from the population of interest. Specifically, the proportion treated in the sample is fixed due to the exposure-stratified sampling, and the distribution of the matched covariates is forced to be the same for the treated and control subjects due to the matching. Although the implications for causal inference are different, this set-up is conceptually similiar to that of a case-control study, where sampling is stratified by outcome (Breslow et al., 2000). As in case-control studies, although the observations in a matched cohort study are not an independent and identically distributed sample from the population distribution of interest, they can be viewed as an independent and identically distributed sample from a particular modified distribution. This is called the biased sampling model framework (Jewell, 1985; Bickel et al., 1993). In a matched cohort study the observations $(Z_1, \ldots, Z_n)$ arise from a biased distribution Q with density

$$q(z) = p(y \mid l, a)p(\bar{w} \mid w, a)p(w \mid a = 1)q(a), \tag{1}$$

and P denotes the distribution of Z in a larger population of interest, with density given by $p(z) = p(y \mid l, a)p(l \mid a)p(a)$ with respect to some dominating measure, and $q(a)$ is the proportion of subjects in the sample receiving treatment level a. In general we write the density under distribution F of variable X evaluated at value c as $f(x = c)$, except when the density we are referring to is unambiguous, e.g., $f(x)$ denotes the density of X under F. The likelihood can be written as $\prod_i p(y_i \mid l_i, a_i)p(\bar{w}_i \mid w_i, a_i)q(a_i) \prod_j p(w_j \mid a = 1)$, where i references units and j references matched

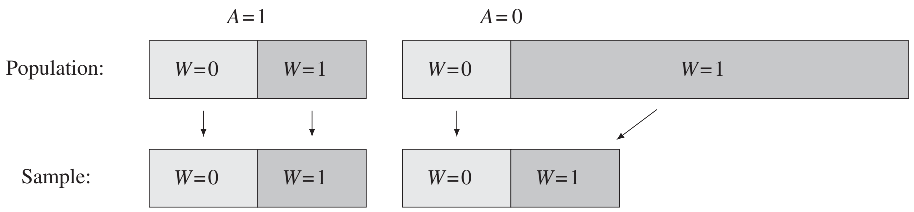

Fig. 1. Schematic of matched cohort study design for 1:1 matching on a binary variable W. Arrows denote random samples of size $n_1/2$.

strata. For illustration Fig. 1 gives a schematic of a matched cohort study in the simple case of 1:1 matching on a binary variable.

In subsequent sections we characterize causal treatment effects using potential outcome notation (Rubin, 1974), and so let Y^a denote the potential outcome that would have been observed had treatment level a been applied. We further make use of some simplifying notation. Specifically we use $\pi(l)$ to denote the propensity score under P given by $p(a=1 \mid l)$, and we use $\xi(l)$ to denote the analog of the propensity score in the biased distribution Q given by $q(a=1 \mid l)$. We also use $\mu(l, a)$ to denote the conditional mean of the outcome given covariates and treatment $E(Y \mid L=l, A=a)$, which is the same under both P and Q whenever it exists. All expectations are taken under the distribution P of interest, unless otherwise noted with a subscript, as in E_Q.

3. Identification and estimation

Throughout we consider the following identifying assumptions, the third of which is commonly called no unmeasured confounding.

Assumption 1 (*Consistency*). If $A=a$ then $Y=Y^a$ with probability one.

Assumption 2 (*Positivity*). For all l such that $p(l)>0$, we have $0<\pi(l)<1$.

Assumption 3 (*Ignorability*). For $a \in \{0,1\}$, $E(Y^a \mid L, A=1) = E(Y^a \mid L, A=0)$.

These assumptions are all typically satisfied by design in randomized trials, but in observational studies they may be violated and are generally untestable. Consistency ensures that one potential outcome is observed for every subject, namely the potential outcome under the treatment that was actually received; it can fail to hold if different versions of treatment have different effects, or if there is interference, for example. Positivity says that treatment is not assigned deterministically, in the sense that every subject has some positive probability of receiving both treatment and control, regardless of covariates. Ignorability says that the mean potential outcomes are the same for both treatment groups once we condition on the covariates, and requires sufficiently many relevant covariates to be collected.

It is well-known and straightforward to show that $E(Y^a) = \int \mu(l,a)p(l)\,\mathrm{d}\nu(l)$ under Assumptions 1–3, where ν is a dominating measure for the distribution of L. Importantly, this expression is identified under P, but not under Q since we observe $q(l) \neq p(l)$ under Q. Note that

$$p(l) = q(\bar{w} \mid w, a=0)p(w \mid a=0)p(a=0) + q(\bar{w} \mid w, a=1)q(w \mid a=1)p(a=1)$$

since $q(\bar{w} \mid w, a) = p(\bar{w} \mid w, a)$ and $q(w \mid a=1) = p(w \mid a=1)$, but at least $p(a)$ is not identified under Q. Without matching, the covariate distributions given treatment $p(l \mid a)$ would be identified, but matching further removes identification of the covariate distribution among the controls since it forces $q(w \mid a=0) = p(w \mid a=1)$. Thus, identification of average effects $E(Y^a)$ cannot be achieved under matched cohort sampling without external knowledge of the treatment proportions $p(a)$ and the matched covariate density $p(w \mid a=0)$.

If $p(a)$ and $p(w \mid a=0)$ are known from external data, however, one can construct estimators of $E(Y^a)$, or any other parameter defined on P, based on appropriately weighted estimating functions, as in van der

Laan et al. (2013). Weighting is necessary since estimating functions based on P will in general be biased, e.g., not have mean zero, under Q. For use in matched cohort studies, estimating functions under P should be weighted by $b(W, A) = \{p(A)/q(A)\}\{p(W \mid A)/p(W \mid a = 1)\}$ since $p(z) = q(z)b(w, a)$.

In many cases such external information is not available, especially when W is high-dimensional. But this is not problematic for estimation of the effect on the treated, which is given by $\psi = E(Y^1 - Y^0 \mid A = 1)$. Under Assumptions 1–3 we have

$$\psi = \int y\, p(y \mid a = 1)\, \mathrm{d}\eta(y) - \int \mu(l, 0)\, p(l \mid a = 1)\, \mathrm{d}\nu(l),$$

where η is a dominating measure for the outcome distribution; this follows from the same logic as in Hahn (1998) and elsewhere. Thus ψ is identified under Assumptions 1–3 in any study design that identifies $p(y \mid l, a)$ and $p(l \mid a = 1)$. Since these densities are components of the density of distribution Q given in (1), it follows that ψ is identified under matched cohort sampling.

As discussed by Breslow et al. (2000) in the context of case-control studies, this fact alone also implies that influence functions for ψ under sampling from Q are equivalent to those under sampling from P, but with densities under distribution Q replacing those under P. For the sake of completeness, we follow Breslow et al. (2000) and prove this result explicitly in the Supplementary Material. To do so we use the same approach as Hahn (1998), with theory developed by Robins & Rotnitzky (1995) and Robins et al. (1995) and discussed in more detail elsewhere (Bickel et al., 1993; van der Laan & Robins, 2003; Tsiatis, 2006). The result can also be derived by weighting the efficient influence function under P by the term $b(W, A)$ as discussed above.

Theorem 1. *The efficient influence function for the effect on the treated ψ under a nonparametric model with distribution Q is*

$$\varphi(\mu, \xi; \psi) = \frac{A}{q(a=1)}\left\{Y - \mu(L, 0) - \psi\right\} - \frac{1-A}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\left\{Y - \mu(L, 0)\right\}.$$

A simple estimator based on the efficient influence function can be formulated by using the efficient influence function φ as an estimating function, after inserting estimates $\hat{\mu}$ and $\hat{\xi}$ of the nuisance functions, i.e., solving $\mathbb{Q}_n\{\varphi(\hat{\mu}, \hat{\xi}; \psi)\} = 0$ where $\mathbb{Q}_n$ is the empirical measure under Q. For example, $\hat{\mu}(l, 0)$ could be predicted values from a regression of the outcome on covariates using only control subjects, and $\hat{\xi}(l)$ could be predicted values from a logistic regression of treatment on covariates. We show that this estimator is doubly robust and derive its asymptotic properties in the Supplementary Material.

Computationally, such estimators are exactly equivalent to those that would be used in a simple study with standard random sampling. Thus, just as in case-control studies where one can ignore the outcome-dependent sampling and regress outcome on exposure using logistic regression to obtain valid odds ratio estimates, Theorem 1 justifies using standard estimators of effects on the treated in cohort studies with exposure-dependent sampling and matching. In particular, one can use propensity score-based estimators as usual even though the propensity score $\pi(l)$ is not identified under matched cohort sampling. In the Supplementary Material we discuss estimation of effect modification among the treated.

4. Efficiency and design

The semiparametric efficiency bound under sampling from Q is the variance of the efficient influence function from Theorem 1. Letting $\sigma^2(l, a) = \mathrm{var}(Y \mid L = l, A = a)$, it is shown in the Supplementary Material that this efficiency bound can be expressed as

$$B_Q = \frac{\Omega + \Sigma_1}{q(a=1)} + \frac{p(a=0)}{p(a=1)}\frac{\Sigma_0^*}{q(a=0)}, \tag{2}$$

where $\Omega = \mathrm{var}\{\mu(L, 1) - \mu(L, 0) \mid A = 1\}$, $\Sigma_1 = E\{\sigma^2(L, 1) \mid A = 1\}$, and $\Sigma_0^* = E\{\varsigma(W) \mid A = 0\}$ with $\varsigma(w) = E[\sigma^2(L, 0)\pi(L)/\{1 - \pi(L)\} \mid W = w, A = 1]$. Letting $\Sigma_0 = E[\sigma^2(L, 0)\pi(L)/\{1 - \pi(L)\} \mid A = 1] = E\{\varsigma(W) \mid A = 1\}$, the efficiency bound under P can be similarly expressed as $B_P = (\Omega + \Sigma_1 + \Sigma_0)/p(a = 1)$.

The expressions for the bounds B_Q and B_P can simplify in certain cases; we will consider three such settings here. The simplest is one in which there are no covariates, i.e., $L = \emptyset$. Then $\pi(l) = p(a = 1)$ so that $\Sigma_0^* = \Sigma_0 = \mathrm{var}(Y \mid A = 0)p(a = 1)/p(a = 0)$, and it also follows that $\Omega = 0$. Another setting of interest is when there are no matching variables, i.e., $W = \emptyset$. Then we again have $\Sigma_0^* = \Sigma_0$, but without further simplification. Lastly we also consider full matching, i.e., $W = L$. Then we have $\Sigma_0^* = \Sigma_0^r$, where $\Sigma_0^r = E\{\sigma^2(L, 0)p(a = 1)/p(a = 0) \mid A = 1\}$ is the value of Σ_0 we would see in a study had all subjects been randomized to treatment with probability $p(a = 1)$ regardless of covariates.

Using the above expressions for B_Q and B_P, it follows that $B_Q < B_P$ if and only if

$$\Sigma_0^* < \frac{q(a = 0)}{p(a = 0)} \left\{ \Sigma_0 - \frac{p(a = 1) - q(a = 1)}{q(a = 1)} \Big(\Omega + \Sigma_1 \Big) \right\}.$$

Clearly, there always exists a cohort study that can match the efficiency bound under random sampling, since random sampling is equivalent to a cohort study with no matching and with $q(a = 1) = p(a = 1)$. In the next theorem we show the more interesting result that there almost always exists a cohort study that is strictly more efficient than random sampling.

THEOREM 2. *Suppose that* $p(a = 1) \neq (\Omega + \Sigma_1)/(\Omega + \Sigma_1 + \Sigma_0)$. *Then there exists a cohort study that is more efficient than random sampling for estimation of* ψ. *For example, an efficiency bound strictly smaller than* B_P *can be attained via an unmatched cohort study with*

$$\min\left\{ p(a = 1), \frac{\Omega + \Sigma_1}{\Omega + \Sigma_1 + \Sigma_0} \right\} < q(a = 1) < \max\left\{ p(a = 1), \frac{\Omega + \Sigma_1}{\Omega + \Sigma_1 + \Sigma_0} \right\}.$$

A proof is given in the Supplementary Material. To illustrate, consider a simple cohort study with no covariates and let $\sigma_a^2 = \mathrm{var}(Y \mid A = a)$. Then any cohort study with

$$p(a = 1) < q(a = 1) < \frac{p(a = 0)\sigma_1^2}{p(a = 0)\sigma_1^2 + p(a = 1)\sigma_0^2},$$

or the inequalities reversed, yields a smaller efficiency bound than random sampling. If treatment is very rare or very common then nearly any cohort study will be more efficient than random sampling, since then the condition approximates $0 < q(a = 1) < 1$.

Matching can provide even more opportunities for efficiency gains. Consider two cohort studies, one without matching, i.e., $W = \emptyset$, yielding efficiency bound B_Q^u and the other fully matched, i.e., $W = L$, yielding efficiency bound B_Q^m. The difference between efficiency bounds then equals

$$B_Q^u - B_Q^m = \frac{1}{q(a = 0)} \frac{p(a = 0)}{p(a = 1)} E\left[\sigma^2(L, 0) \left\{ \frac{\pi(L)}{1 - \pi(L)} - \frac{p(a = 1)}{p(a = 0)} \right\} \Bigm| A = 1 \right].$$

If there is no confounding so that $\pi(l) = p(a = 1)$, then the bounds are clearly equal and matching does not provide any efficiency gains. However, when there is confounding the above will often be positive since $\pi(l)$ will generally be larger than $p(a = 1)$ among the treated. For example, if $\sigma^2(l, 0)$ is constant then $B_Q^u \geqslant B_q^m$ by Jensen's inequality. This suggests that matched cohort studies will in general provide better efficiency than unmatched cohort studies.

In principle one could design a fully efficient matched cohort study by minimizing the expression for B_Q given in (2) over choices of $q(a = 1) = 1/(k + 1)$ and different sets of matching variables W. Optimizing over different matching variables would be difficult in practice, but results for optimizing over $q(a = 1)$ are given in the following theorem.

Table 1. *Bias, variance, and coverage based on* 500 *simulated* 1 : 1 *matched cohort studies*

		Correct model							
Sample		Neither		Treatment		Outcome		Both	
size	Estimator	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov
100	IP-weighted	−20 (129)	98	6 (172)	97	−20 (129)	98	6 (172)	97
	Regression	−54 (29)	68	−54 (29)	68	0 (2·3)	95	0 (2·3)	95
	Doubly robust	−41 (36)	76	−7 (35)	94	0 (2·5)	94	0 (2·7)	92
1000	IP-weighted	−27 (83)	84	−1 (95)	96	−27 (83)	84	−1 (95)	96
	Regression	−55 (30)	0	−55 (30)	0	0 (2·2)	95	0 (2·2)	95
	Doubly robust	−41 (33)	4	−1 (30)	94	0 (2·4)	95	0 (2·5)	96

IP, inverse-probability; SE, empirical standard error multiplied by $n^{1/2}$; Cov, coverage (%).

Theorem 3. *Consider a cohort study with a fixed set of, possibly empty, matching variables and given sample size. The optimal number of matches that maximizes efficiency for estimation of* ψ *is* $k_{opt} = [\{p(a=0)/p(a=1)\}\{\Sigma_0^*/(\Omega+\Sigma_1)\}]^{1/2}$.

In the simplest matched cohort study with no covariates, this expression simplifies to $k_{opt} = \sigma_0/\sigma_1$. Thus for such studies the optimal matching ratio does not depend on the treatment prevalence, and in particular 1:1 matching is optimal if the variance of the outcome is constant across treatment groups. As intuition would suggest, if the variance of the outcome is greater among controls then more matched controls should be used, and if the variance is greater among the treated then fewer matched controls should be used.

5. Simulations and illustration

5·1. *Simulation study*

To explore finite-sample properties we adapt the simulation set-up from Kang & Schafer (2007). Specifically we simulated $L_j \sim N(0, 1)$ for $j = 1, \ldots, 4$, $\pi(l) = \text{expit}(-1{\cdot}7 - l_1 + 0{\cdot}5l_2 - 0{\cdot}25l_3 - 0{\cdot}1l_4)$ so that $p(a=1) = 0{\cdot}20$, and $Y = \mu(L, A) + \epsilon$ for $\mu(l, a) = 200 + 13{\cdot}7l_1 + 13{\cdot}7\sum_j l_j + 10a$, and $\epsilon \sim N(0, 1)$ so that $\psi = 10$. We generated matched cohort studies with $q(a=1) = 0{\cdot}5$ and $W = I(L_1 > 0)$, which ensures that $q(a = 1 \mid l)$ follows a logistic model with covariates w and l. For each simulated dataset we applied inverse-probability-weighted, regression, and doubly robust estimators, with confidence intervals computed via sandwich standard errors. To misspecify models we transformed L as in Kang & Schafer (2007).

As shown in Table 1, the inverse-probability-weighted and regression estimators were biased when relying on misspecified models, while the doubly robust estimator performed well as long as at least one model was correct. The doubly robust and regression estimators had similar efficiency when the outcome model was correct; when only the treatment model was correct the doubly robust estimator was more efficient than the inverse-probability-weighted estimator. Coverage was near 95% except under misspecification. In the Supplementary Material we give further results comparing with random sampling and different matching ratios.

5·2. *Application*

Here we analyse the 3:1 matched cohort study by Ingelsson et al. (2011) discussed in § 1. We used the same three estimators as in the simulation study, with logistic regression models for the treatment, i.e., hysterectomy, and outcome, i.e., cardiovascular disease within 10 years after enrolment. The matching covariates were year of birth and county of residence, and the unmatched covariates were socioeconomic status and age at enrolment. For simplicity we assumed independent censoring. As shown in Table 2, assuming no unmeasured confounding we estimate that hysterectomy yielded a statistically significant 0·55% increased risk of cardiovascular disease within 10 years, among those who underwent hysterectomy.

We also used the formulas from § 4 to analyse efficiency, by estimating the terms in the bound B_Q. For simplicity we assumed $p(w \mid a) = p(w)$ and focused on varying $p(a)$. We estimate that 3:1 matched

Table 2. *Hysterectomy and* 10*-year cardiovascular risk*

Method	Estimate (%)	SE (%)	95% CI	p-value
IP-weighted	0·47	0·093	(0·29, 0·65)	$< 0{\cdot}001$
Regression	0·55	0·092	(0·37, 0·73)	$< 0{\cdot}001$
Doubly robust	0·55	0·092	(0·37, 0·73)	$< 0{\cdot}001$

CI, confidence interval; SE, standard error; IP, inverse-probability.

cohort sampling yields a smaller efficiency bound than random sampling if $p(a = 1) < 23\%$, and is more than twice as efficient if $p(a = 1) < 7\%$. We also estimate that 3:1 matching is optimal if $p(a = 1) = 3\%$, and that full matching using socioeconomic status and age is beneficial if $p(a = 1) < 23\%$. More details are in the Supplementary Material.

Acknowledgement

We thank two reviewers and an associate editor for helpful comments. Edward Kennedy acknowledges support from the U.S. National Institutes of Health, Arvid Sjölander from the Swedish Research Council, and Dylan Small from the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs, more details about the simulation and illustration, and an R function for implementing the estimators.

References

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–67.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Bickel, P. J., Klaassen, C. A., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Breslow, N. E., Robins, J. M. & Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**, 447–55.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–31.

Heckman, J. J. & Todd, P. E. (2009). A note on adapting propensity score matching and selection models to choice based samples. *Economet. J.* **12**, S230–4.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Studies* **64**, 605–54.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Rev. Econ. Studies* **65**, 261–94.

Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–89.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Statist.* **86**, 4–29.

Ingelsson, E., Lundholm, C., Johansson, A. L. & Altman, D. (2011). Hysterectomy and risk of cardiovascular disease: A population-based cohort study. *Eur. Heart J.* **32**, 745–50.

Jewell, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* **72**, 11–21.

Jewell, N. P. (2003). *Statistics for Epidemiology*. London: CRC Press.

Kang, J. D. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–39.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *Am. Econ. Rev.* **101**, 532–7.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.* **90**, 122–9.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.

Rothman, K. J., Greenland, S. & Lash, T. L. (2008). *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688.

Sjölander, A. & Greenland, S. (2013). Ignoring the matching variables in cohort studies–when is it valid and why? *Statist. Med.* **32**, 4696–708.

Sjölander, A., Johansson, A. L., Lundholm, C., Altman, D., Almqvist, C. & Pawitan, Y. (2012). Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Statist. Sci.* **27**, 395–411.

Tchetgen Tchetgen, E. J. & Rotnitzky, A. (2011). Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. *Statist. Med.* **30**, 335–47.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

van der Laan, M. J., Petersen, M. & Zheng, W. (2013). Estimating the effect of a community-based intervention with two communities. *J. Causal Infer.* **1**, 83–106.

[*Received October* 2014. *Revised March* 2015]

Biometrika (20xx), **99**, 1, *pp.* 1–10
© 20xx Biometrika Trust
Printed in Great Britain

Advance Access publication on 31 July 20xx

Supplementary materials for "Semiparametric causal inference in matched cohort studies"

BY E.H. KENNEDY

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, U.S.A.

kennedye@mail.med.upenn.edu

A. SJÖLANDER

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden

arvid.sjolander@ki.se

AND D.S. SMALL

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, U.S.A.

dsmall@wharton.upenn.edu

1. BRIEF LITERATURE REVIEW

There are many important papers on semiparametric estimation of the effect on the treated in a simple random sampling setting. Here we give a brief description of this literature. Rubin (1977) and Heckman & Robb (1985) were two of the earlier papers to discuss effects on the treated in some detail. Later, Heckman et al. (1997) and Heckman et al. (1998) considered kernel-based matching approaches for estimation, including using an estimated propensity score. Hahn (1998) derived the efficient influence function for the effect on the treated (under a nonparametric model and a model in which the propensity score is known), and developed semiparametric efficient estimators that rely on nonparametric estimation of the propensity score and outcome regression functions. Dehejia & Wahba (1999) used stratification and matching on the propensity score with the data from LaLonde (1986), and found that the estimates were more similar to a randomized trial benchmark than those based on regression. Hirano & Imbens (2001) considered doubly robust estimation based on the efficient influence function, and used the approach to estimate effects of right heart catheterization. Hirano et al. (2003) discussed a potentially efficient estimator that only relies on estimation of the propensity score. Imbens (2004) gave a broad overview of semiparametric methods for estimating treatment effects. Abadie & Imbens (2006) derived asymptotic theory for matching estimators that use a fixed number of matches, and showed in Abadie & Imbens (2008) that the standard bootstrap is generally not valid for such estimators. More recently, Chen et al. (2008) generalized much of the above work to settings involving non-linear, possibly non-smooth, over-identified moment conditions, and considered a general context in which results can be applied to missing data and measurement error problems as well as causal inference. Kline (2011) showed that an early estimator of the effect on the treated, proposed by Oaxaca (1973) and Blinder (1973), actually fits in the doubly robust framework proposed by Robins et al. (1994) and further developed elsewhere, and gave a re-analysis of the LaLonde (1986) data. Zhang et al. (2012) considered quantile effects on the treated.

2. EFFECT MODIFICATION

In many studies interest centers not just on marginal treatment effects, but also on how effects can change with covariates. The average effect on the treated conditional on putative effect modifiers $V \subseteq L$ is given by $\gamma(v) = E(Y^1 - Y^0 \mid V = v, A = 1)$, and is identified under Assumptions 1-3 in the main text by the expression

$$\gamma(v) = \int y \, p(y \mid v, a = 1) \; d\eta(y) - \int \mu(l, 0) p(\overline{v} \mid v, a = 1) \; d\nu(l).$$

Thus the conditional effect $\gamma(v)$ is identified in any study design that identifies $p(y \mid l, a)$ and $p(\overline{v} \mid v, a = 1)$, including matched cohort studies.

In the next section we derive the efficient influence function for $\gamma(v)$ under a nonparametric model with distribution Q. However, in practice V might include variables with many levels or continuous components, so that specifying a saturated model is impossible. In such cases we might want to assume a parsimonious model $\gamma(v; \psi)$ for $\gamma(v)$, which is indexed by finite-dimensional parameter $\psi \in \mathbb{R}^p$. One approach in this setting is to develop estimators under the assumption that this model is exactly correctly specified. See Lei (2011) for nice work in this setting, which can yield important efficiency advantages when the effect modification can be modeled well. An alternative approach is to only assume $\gamma(v; \psi)$ is a possibly misspecified working model and define the target parameter of interest as a projection of $\gamma(v)$ onto the working model (Neugebauer & van der Laan, 2007). We take the latter approach, defining ψ as the minimizer of the distance $\int \{\gamma(v) - \gamma(v; \psi)\}^2 p(v \mid a = 1) \; d\nu(v)$. If the working model is incorrect, this parameter is still validly defined as a projection. Further, if the working model happens to be correct, the efficient influence function for ψ derived under the working model assumption will still be valid under the assumption that the model is correct, just not necessarily efficient.

THEOREM 4. *Let $\gamma(v; \psi)$ be a working model for $\gamma(v)$ with $\psi \in \mathbb{R}^p$, so that ψ is defined as the projection* $\arg\min_{\psi^* \in \mathbb{R}^p} \int \{\gamma(v) - \gamma(v; \psi^*)\}^2 p(v \mid a = 1) \; d\nu(v)$*. The efficient influence function for ψ under a nonparametric model with distribution Q is then $-E[\{\partial\gamma(\mu, \xi; \psi)/\partial\psi\}^{\otimes 2} \mid A = 1]^{-1} \varphi^*(\mu, \xi; \psi)$, where $\varphi^*(\mu, \xi; \psi)$ is defined as*

$$\frac{\partial \gamma(V; \psi)}{\partial \psi} \left[\frac{A}{q(a = 1)} \Big\{ Y - \mu(L, 0) - \gamma(V; \psi) \Big\} - \frac{1 - A}{q(a = 1)} \left\{ \frac{\xi(L)}{1 - \xi(L)} \right\} \Big\{ Y - \mu(L, 0) \Big\} \right].$$

3. PROOFS FOR SECTION 3 IN MAIN TEXT

Here we prove results for $\gamma(v)$, since results for $E(Y^1 - Y^0 \mid A = 1)$ follow by taking $V = \emptyset$. We write expectations of g under F as either $E_F(g)$ or $Fg = \int g \; dF$, but use $E_P = E$.

Proof (Identification). It follows that $E(Y^1 \mid V = v, A = 1) = \int y \;\; p(y \mid v, a = 1) \;\; d\eta(y)$ from the consistency assumption alone. Then

$$\begin{aligned} E(Y^0 \mid V = v, A = 1) &= \int E(Y^0 \mid L = l, A = 1) \, p(\overline{v} \mid v, a = 1) \; d\nu(\overline{v}) \\ &= \int E(Y^0 \mid L = l, A = 0) \, p(\overline{v} \mid v, a = 1) \; d\nu(\overline{v}) = \int \mu(l, 0) \, p(\overline{v} \mid v, a = 1) \; d\nu(\overline{v}) \end{aligned}$$

where the first equality follows by iterated expectation, the second by ignorability, and the third by consistency. Positivity is required so as to prevent conditioning on null sets. □

Proof (Theorems 1 and 4). Let $q(z;\epsilon)$ be a parametric submodel with parameter $\epsilon \in \mathbb{R}$ and $q(z;0) = q(z)$. Recalling the identifying expression of $\gamma(v)$, we write

$$\gamma(v;\epsilon) = \int\int y\Big\{p(y \mid l, a = 1;\epsilon) - p(y \mid l, a = 0;\epsilon)\Big\}p(\overline{v} \mid v, a = 1;\epsilon)\; d\eta(y) d\nu(\overline{v}).$$

By definition the efficient influence function under Q is the unique function $\varphi(Z)$ that satisfies $\partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0} = E_Q\{\varphi(Z)S_\epsilon(Z)\}$, where $S_\epsilon(Z)$ is defined as $\partial \log q(z;\epsilon)/\partial\epsilon|_{\epsilon=0}$ with

$$\frac{\partial \log q(z;\epsilon)}{\partial\epsilon}\Big|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}\Big\{\log p(y \mid l, a;\epsilon) + \log p(\overline{v} \mid v, a;\epsilon) + \log p(v \mid a = 1;\epsilon) + \log q(a;\epsilon)\Big\}\Big|_{\epsilon=0}.$$

We denote the four terms on the right as $S_y(y, l, a)$, $S_{\overline{v}}(\overline{v}, v, a)$, $S_v(v)$, and $S_a(a)$. Then it is straightforward to show that $\partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0}$ equals

$$E\bigg(Y\Big\{S_y(Y, L, 1) + S_{\overline{v}}(\overline{V}, V, 1)\Big\} - E\Big[Y\Big\{S_y(Y, L, 0) + S_{\overline{v}}(\overline{V}, V, 1)\Big\} \;\Big|\; L, A = 0\Big] \;\Big|\; V = v, A = 1\bigg).$$

Denote the putative efficient influence function as

$$\varphi(Z) = \frac{I(V = v)}{p(v \mid a = 1)}\left[\frac{A}{q(a = 1)}\Big\{Y - \mu(L, 0) - \gamma(v)\Big\} - \frac{1 - A}{q(a = 1)}\left\{\frac{\xi(L)}{1 - \xi(L)}\right\}\Big\{Y - \mu(L, 0)\Big\}\right].$$

Then one can also verify that $E_Q\{\varphi(Z)S_\epsilon(Z)\} = \partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0} - h(v)$ where $h(v)$ equals

$$\begin{aligned}
&E\left[\frac{q(a = 0)}{q(a = 1)}\left\{\frac{\xi(L)}{1 - \xi(L)}\right\}\Big\{Y - \mu(L, 0)\Big\}\Big\{S_{\overline{v}}(\overline{V}, V, 0) + S_v(V) + S_a(0)\Big\} \;\Big|\; V = v, A = 0\right] \\
&\quad + E\left[\Big\{\mu(L, 0) + \gamma(v)\Big\}S_y(Y, L, 1) \;\Big|\; V = v, A = 1\right] + E\left\{\gamma(v)S_{\overline{v}}(\overline{V}, V, 1) \;\Big|\; V = v, A = 1\right\} \\
&\quad - E\left[\frac{q(a = 0)}{q(a = 1)}\left\{\frac{\xi(L)}{1 - \xi(L)}\right\}\mu(L, 0)S_y(Y, L, 0) \;\Big|\; V = v, A = 0\right] \\
&\quad - E\left[\Big\{Y - \mu(L, 0) - \gamma(v)\Big\}\Big\{S_v(V) + S_a(1)\Big\} \;\Big|\; V = v, A = 1\right].
\end{aligned}$$

However, the first line above is zero by iterated expectations since $E(Y \mid L, A = 0) - \mu(L, 0) = 0$. The second and third lines are zero by iterated expectations and standard properties of conditional score functions, in particular that $E\{S_y(Y, L, A) \mid L, A\} = E\{S_{\overline{v}}(\overline{V}, V, A) \mid V, A\} = 0$. Similarly the fourth line is zero by the definition of $\gamma(v)$. Therefore $E_Q\{\varphi(Z)S_\epsilon(Z)\} = \partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0}$ and it follows that $\varphi(Z)$ is the efficient influence function. □

Proof (Double robustness). Here we show that $E_Q\{\varphi^*(\mu,\xi;\psi_0)\}=0$ if $\tilde{\mu}=\mu$ or $\tilde{\xi}=\xi$ (not necessarily both). Note that we can write this expectation as

$$\begin{aligned}
&E_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\frac{A}{q(a=1)}\Big\{Y-\tilde{\mu}(L,0)-\gamma(V;\psi)\Big\}-\frac{1-A}{q(a=1)}\left\{\frac{\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{Y-\tilde{\mu}(L,0)\Big\}\bigg]\bigg)\\
&=E_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\frac{\xi(L)}{q(a=1)}\{\mu(L,1)-\tilde{\mu}(L,0)-\gamma(V;\psi)\}\\
&\qquad\qquad-\frac{1-\xi(L)}{q(a=1)}\left\{\frac{\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\{\mu(L,0)-\tilde{\mu}(L,0)\}\bigg]\bigg)\\
&=\frac{1}{q(a=1)}E_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\xi(L)\Big\{\mu(L,1)-\mu(L,0)-\gamma(V;\psi)\Big\}\\
&\qquad\qquad-\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\bigg)\\
&=\int\frac{\partial\gamma(v;\psi)}{\partial\psi}\Big\{\mu(l,1)-\mu(l,0)-\gamma(v;\psi)\Big\}p(l\mid a=1)\,d\nu(l)\\
&\qquad\qquad-\frac{1}{q(a=1)}E_Q\bigg[\frac{\partial\gamma(V;\psi)}{\partial\psi}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\\
&=\int\frac{\partial\gamma(v;\psi)}{\partial\psi}\Big\{\gamma(v)-\gamma(v;\psi)\Big\}p(v\mid a=1)\,d\nu(v)\\
&\qquad\qquad-\frac{1}{q(a=1)}E_Q\bigg[\frac{\partial\gamma(V;\psi)}{\partial\psi}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\\
&=\frac{1}{q(a=1)}E_Q\left[\frac{\partial\gamma(V;\psi)}{\partial\psi}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\right].
\end{aligned}$$

The first equality follows by definition, the second by iterated expectation, the third by adding and subtracting $\mu(L,0)$ and rearranging, the fourth since

$$\int\frac{\xi(l)}{q(a=1)}g(l)q(l)\,d\nu(l)=\int g(l)\frac{q(a=1\mid l)q(l)}{q(a=1)}\,d\nu(l)=\int g(l)q(l\mid a=1)\,d\nu(l)$$

and $q(l\mid a=1)=p(l\mid a=1)$, the fifth by iterated expectation, and the last by the fact that $\int\partial\gamma(v;\psi)/\partial\psi\{\gamma(v)-\gamma(v;\psi)\}p(v\mid a=1)\ \ d\nu(v)=0$ by definition when $\psi=\arg\min_{\psi^*\in\mathbb{R}^p}\int\{\gamma(v)-\gamma(v;\psi^*)\}^2p(v\mid a=1)\,d\nu(v)$.

Now the result follows since the term after the last equality reduces to zero whenever either $\tilde{\mu}=\mu$ or $\tilde{\xi}=\xi$. □

Proof (Asymptotic normality). Define $\hat{\psi}$ as the solution to $\mathbb{Q}_n\varphi(\psi,\hat{\eta})=0$ where $\eta=(\mu,\xi)$, let $||f||^2=\int f^2dQ$ denote the squared $L_2(Q)$ norm, and assume

1. $\hat{\psi}-\psi_0=o_p(1)$, $||\hat{\mu}-\tilde{\mu}||=o_p(1)$, and $||\hat{\xi}-\tilde{\xi}||=o_p(1)$ with either $\tilde{\mu}=\mu_0$ or $\tilde{\xi}=\xi_0$.
2. $\varphi(\psi,\hat{\eta})$ lies in a Donsker class with probability one as $n\to\infty$.
3. The map $\psi\to Q\varphi(\psi,\eta)$ is differentiable at ψ_0 uniformly in η, with derivative $D_{\psi,\eta}$.
4. $\varphi(\psi,\eta)$ is continuous in $L_2(Q)$ at $(\psi_0,\tilde{\eta})$.

We will show that if $\tilde\eta = \eta_0$ and $||\hat\mu - \mu_0|| \cdot ||\hat\xi - \xi_0|| = o_p(n^{-1/2})$ then $\hat\psi$ is root-n consistent, asymptotically normal, and efficient. If $\eta \in \mathbb{R}^d$, the map $\eta \to Q\varphi(\psi, \eta)$ is differentiable with nonsingular derivative $\Delta_{\psi,\eta}$, and $\hat\eta$ has influence function $\phi(\eta)$ so that $\hat\eta - \tilde\eta = \mathbb{Q}_n\phi(\tilde\eta) + o_p(n^{-1/2})$, then $\hat\psi$ is root-n consistent and asymptotically normal, even if $\tilde\eta \neq \eta_0$ (i.e., even if one of $\tilde\mu \neq \mu_0$ or $\tilde\xi \neq \xi_0$).

By Theorem 5.31 from van der Vaart (2000) (also see van der Vaart (2002)), under Assumptions 1-4 above we have

$$\hat\psi - \psi_0 = -D^{-1}_{\psi_0,\tilde\eta}\, Q\varphi(\psi_0,\hat\eta) - D^{-1}_{\psi_0,\tilde\eta}\, \mathbb{Q}_n\varphi(\psi_0,\tilde\eta) + o_p\Big(n^{-1/2} + ||Q\varphi(\psi_0,\hat\eta)||\Big).$$

Further from the double robustness result on the previous page we have

$$Q\varphi(\psi_0,\hat\eta) = \frac{1}{q(a=1)} Q\left[\frac{\partial\gamma(V;\psi_0)}{\partial\psi}\Big\{\mu_0(L,0) - \hat\mu(L,0)\Big\}\left\{\frac{\xi_0(L) - \hat\xi(L)}{1 - \hat\xi(L)}\right\}\right].$$

First assume $\tilde\eta = \eta_0$ and $||\hat\mu - \mu_0|| \cdot ||\hat\xi - \xi_0|| = o_p(n^{-1/2})$. Since $Q(fg) \leq ||f|| \cdot ||g||$ by the Cauchy-Schwarz inequality, for some constant C it follows that

$$Q\varphi(\psi_0,\hat\eta) \leq C||\hat\mu - \mu_0|| \cdot ||\hat\xi - \xi_0||,$$

and the right-hand side is $o_p(n^{-1/2})$ by assumption. Therefore

$$\hat\psi - \psi_0 = -D^{-1}_{\psi_0,\eta_0}\mathbb{Q}_n\varphi(\psi_0,\eta_0) + o_p(n^{-1/2})$$

so that $\hat\psi$ is root-n consistent, asymptotically normal, and efficient.

Now assume $\eta \in \mathbb{R}^d$, the map $\eta \to Q\varphi(\psi,\eta)$ is differentiable, and $\hat\eta$ has influence function $\phi(\eta)$. Then by the Delta method we have

$$Q\varphi(\psi_0,\hat\eta) = Q\varphi(\psi_0,\hat\eta) - Q\varphi(\psi_0,\tilde\eta) = \Delta_{\psi_0,\tilde\eta}\, \mathbb{Q}_n\phi(\tilde\eta) + o_p(n^{-1/2}).$$

Therefore $\hat\psi - \psi_0 = -D^{-1}_{\psi_0,\tilde\eta}\ \Delta_{\psi_0,\tilde\eta}\ \mathbb{Q}_n\phi(\tilde\eta) - D^{-1}_{\psi_0,\tilde\eta}\ \mathbb{Q}_n\varphi(\psi_0,\tilde\eta) + o_p(n^{-1/2} + O_p(n^{-1/2}))$, and this implies that

$$\hat\psi - \psi_0 = -D^{-1}_{\psi_0,\tilde\eta}\ \mathbb{Q}_n\Big\{\Delta_{\psi_0,\tilde\eta}\ \phi(\tilde\eta) + \varphi(\psi_0,\tilde\eta)\Big\} + o_p(n^{-1/2}),$$

so that $\hat\psi$ is root-n consistent and asymptotically normal. □

4. Proofs for Section 4 in Main Text

Proof (Efficiency bound). Using notation from the main text, we have

$$\begin{aligned}
&\text{var}_Q\left[\frac{A}{q(a=1)}\Big\{Y - \mu(L,0) - \psi\Big\} - \frac{1-A}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\Big\{Y - \mu(L,0)\Big\}\right] \\
&\quad = \frac{1}{q(a=1)^2}E_Q\Big[\xi(L)\sigma^2(L,1) + \{\mu(L,1) - \mu(L,0) - \psi\}^2\xi(L) + \frac{\xi(L)^2}{1-\xi(L)}\sigma^2(L,0)\Big] \\
&\quad = \frac{1}{q(a=1)}E\Big[\sigma^2(L,1) + \{\mu(L,1) - \mu(L,0) - \psi\}^2 + \frac{\xi(L)}{1-\xi(L)}\sigma^2(L,0)\ \Big|\ A=1\Big] \\
&\quad = \frac{\Omega + \Sigma_1}{q(a=1)} + \frac{1}{q(a=1)}E\left[\frac{\pi(L)}{1-\pi(L)}\frac{q(a=1)}{q(a=0)}\frac{p(a=0)}{p(a=1)}\frac{p(W \mid a=0)}{p(W \mid a=1)}\sigma^2(L,0)\middle| A=1\right],
\end{aligned}$$

and the result follows since $\Sigma_0^* = E\left[\frac{\pi(L)}{1-\pi(L)}\frac{p(W|a=0)}{p(W|a=1)}\sigma^2(L,0) \mid A=1\right]$. □

Proof (Condition for $B_Q < B_P$). Using the expressions for B_Q and B_P, we have

$$\begin{aligned} B_Q < B_P &\iff \frac{\Omega + \Sigma_1}{q(a=1)} + \frac{p(a=0)}{p(a=1)} \frac{\Sigma_0^*}{q(a=0)} < \frac{\Omega + \Sigma_1 + \Sigma_0}{p(a=1)} \\ &\iff \Sigma_0^* < \frac{q(a=0)}{p(a=0)} p(a=1) \left\{ \frac{\Omega + \Sigma_1 + \Sigma_0}{p(a=1)} - \frac{\Omega + \Sigma_1}{q(a=1)} \right\} \\ &\iff \Sigma_0^* < \frac{q(a=0)}{p(a=0)} \left\{ \Sigma_0 - \frac{p(a=1) - q(a=1)}{q(a=1)} \Big(\Omega + \Sigma_1 \Big) \right\}. \end{aligned}$$

This gives the desired result. □

Proof (Theorem 2). If there is no matching then $\Sigma_0^* = \Sigma_0$, and therefore

$$\begin{aligned} B_Q < B_P &\iff \Sigma_0 < \frac{q(a=0)}{p(a=0)} \left\{ \Sigma_0 - \frac{p(a=1) - q(a=1)}{q(a=1)} \Big(\Omega + \Sigma_1 \Big) \right\} \\ &\iff \Big\{ p(a=1) - q(a=1) \Big\} \frac{q(a=0)}{q(a=1)} < \Big\{ p(a=1) - q(a=1) \Big\} \frac{\Sigma_0}{\Omega + \Sigma_1}. \end{aligned}$$

If $p(a=1) > q(a=1)$ then the above is equivalent to $q(a=1) > (\Omega + \Sigma_1)/(\Omega + \Sigma_1 + \Sigma_0)$, while if $p(a=1) < q(a=1)$ then it is equivalent to $q(a=1) < (\Omega + \Sigma_1)/(\Omega + \Sigma_1 + \Sigma_0)$. □

Proof (Theorem 3). Since $q(a=1) = 1/(k+1)$ we can write the efficiency bound as

$$B_Q = (k+1)(\Omega + \Sigma_1) + \left(\frac{k+1}{k} \right) \frac{p(a=0)}{p(a=1)} \Sigma_0^* = kc_1 + \frac{c_2}{k} + (c_1 + c_2),$$

where $c_1 = \Omega + \Sigma_1$ and $c_2 = \{p(a=0)/p(a=1)\}\Sigma_0^*$. We want to find the value of k that minimizes this expression. The derivative with respect to k is $c_1 - (c_2/k^2)$, which when solved for k yields $k^* = (c_2/c_1)^{1/2}$. This is guaranteed to be a minimum since the second derivative at this value is $2c_2/(k^*)^3$, and both c_1 and c_2 (and thus also k^*) are necessarily positive. Therefore $k_{opt} = (c_2/c_1)^{1/2}$. □

5. ADDITIONAL SIMULATION RESULTS

In Table 3 we give additional simulation results comparing matched cohort sampling (with 1:1 and 3:1 matching) versus random sampling, using the same simulation model as described in the main text. For this simulation setting we have $p(a=1) \approx 0.203$ and

$$\Omega = 0, \ \Sigma_1 = 1, \ \text{and} \ \Sigma_0^* \approx 0.495,$$

so that the optimal number of matched controls is approximately 1.4. Therefore 1:1 matching should be more efficient than 3:1 matching and random sampling, at least for the doubly robust estimator under correct model specification; in fact this is exactly what we see.

For our simulations, in general, the estimators applied in 1:1 matched cohort samples were more efficient than in 3:1 matched cohort samples, which were more efficient than in random samples of the same size. However this relation did not always hold when models were misspecified, or for inverse-probability-weighted estimators even under correct model specification. This is to be expected based on theory, since under model misspecification and with inefficient estimators there are generally no theoretical efficiency guarantees.

Table 3. *Percent bias, scaled empirical standard errors, and confidence interval coverage based on 500 simulated datasets*

		Correct model							
	Sampling	Neither		Treatment		Outcome		Both	
n	Estimator	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov
10^2	*MCS (1:1)*								
	IPW	-20 (129)	98	6 (172)	97	-20 (129)	98	6 (172)	97
	Reg	-54 (29)	68	-54 (29)	68	0 (2.3)	95	0 (2.3)	95
	DR	-41 (36)	76	-7 (35)	94	0 (2.5)	94	0 (2.7)	92
	MCS (3:1)								
	IPW	-25 (139)	99	13 (109)	99	-25 (139)	99	13 (109)	99
	Reg	-53 (37)	68	-53 (37)	68	0 (2.5)	94	0 (2.5)	94
	DR	-37 (41)	81	-5 (30)	97	0 (2.7)	95	0 (2.7)	94
	SRS								
	IPW	-20 (197)	99	13 (171)	98	-20 (197)	99	13 (171)	98
	Reg	-97 (53)	58	-97 (53)	58	0 (2.7)	95	0 (2.7)	95
	DR	-44 (81)	81	-10 (59)	96	0 (3.2)	93	0 (3.4)	93
10^3	*MCS (1:1)*								
	IPW	-27 (83)	84	-1 (95)	96	-27 (83)	84	-1 (95)	96
	Reg	-55 (30)	0	-55 (30)	0	0 (2.2)	95	0 (2.2)	95
	DR	-41 (33)	4	-1 (30)	94	0 (2.4)	95	0 (2.5)	96
	MCS (3:1)								
	IPW	-30 (58)	63	1 (62)	98	-30 (58)	63	1 (62)	98
	Reg	-56 (37)	0	-56 (37)	0	0 (2.4)	95	0 (2.4)	95
	DR	-38 (38)	8	-1 (26)	96	0 (2.5)	95	0 (2.6)	95
	SRS								
	IPW	-21 (81)	89	1 (92)	96	-21 (81)	89	1 (92)	96
	Reg	-101 (55)	0	-101 (55)	0	0 (2.7)	95	0 (2.7)	95
	DR	-43 (48)	19	-1 (49)	95	0 (2.9)	94	0 (3.0)	94
10^4	*MCS (1:1)*								
	IPW	-25 (75)	9	0 (88)	96	-25 (75)	9	0 (88)	96
	Reg	-56 (31)	0	-56 (31)	0	0 (2.1)	96	0 (2.1)	96
	DR	-41 (34)	0	0 (30)	96	0 (2.3)	95	0 (2.3)	96
	MCS (3:1)								
	IPW	-30 (58)	0	0 (61)	95	-30 (58)	0	0 (61)	95
	Reg	-56 (36)	0	-56 (36)	0	0 (2.5)	95	0 (2.5)	95
	DR	-38 (37)	0	0 (25)	94	0 (2.6)	94	0 (2.6)	95
	SRS								
	IPW	-20 (75)	23	0 (89)	94	-20 (75)	23	0 (89)	94
	Reg	-102 (58)	0	-102 (58)	0	0 (2.6)	96	0 (2.6)	96
	DR	-43 (49)	0	0 (50)	95	0 (2.7)	97	0 (2.8)	97

SE, standard error multiplied by $n^{1/2}$; IPW, inverse-probability-weighted; Reg, regression; DR, doubly robust. *MCS (k:1)* denotes a matched cohort sampling with k:1 matching on $W \in \{0, 1\}$, and *SRS* denotes simple random sampling.

6. ADDITIONAL ILLUSTRATION DETAILS

In the efficiency analysis given in the main text (but not the main analysis estimating the effect of hysterectomy), we assumed for simplicity that the distribution of the matching covariates was the same for the treated and controls, i.e., $p(w \mid a) = p(w)$. This simplifying assumption allowed us to focus the discussion on how the efficiency bounds compare when varying the marginal proportion treated $p(a = 1)$. However, since most studies match on variables that are thought to be strong confounders, typically we would expect $p(w \mid a = 0)$ to be far from equal to $p(w \mid a = 1)$. Therefore in practice it would often be preferable to specify different values of $p(w \mid a = 0)$, as was done for $p(a = 1)$ in the main text, and see how the bounds under Q and P change relative to each other. Also note that although the bound under P is not identifiable under a matched sampling scheme, the bound under Q is identifiable. In particular, assuming correctly specified treatment and outcome models, it can be estimated with an estimate of the variance (under Q) of the doubly robust estimator.

7. R CODE

```
require(sandwich)

#---INPUT---
#Y: string; name of outcome in data
#Yformula: glm formula for outcome,
#note: this model is fitted to the unexposed
#(i.e. those with A=0), so the formula should not contain A
#Yfamily: glm family for outcome (only used for link function)
#A: string; name of exposure (coded as 0/1) in data
#Aformula: logistic formula for exposure
#method: string; estimation method ("ML", "IPW", "DR", or "DRwt").
#"DRwt" gives DR estimation with weighted LS outcome regression
#cluster: name of cluster id variable
#data: dataset containing all variables

#---OUTPUT---
#psi: estimate of psi
#se: standard error for the estimate of psi

matched <- function(Y,Yformula,Yfamily,A,Aformula,method,cluster,data){

  #preparation
  unexposed <- which(data[,A]==0)
  data0 <- data[unexposed,]; data0star <- data; data0star[,A] <- 0
  A <- data[,A]; Y <- data[,Y]; n <- nrow(data)
  if(missing(cluster)){ ncluster <- n } else {
    ncluster <- length(unique(data[,cluster])) }

  #fit models
  if(method=="IPW" | method=="DR" | method=="DRwt"){
    Afit <- glm(formula=Aformula,family="binomial",data=data)
    w <- exp(predict(object=Afit,newdata=data,type="link"))
    #w = omega/(1-omega)
    data0$w <- w[unexposed]; nApar <- length(Afit$coef)
    LA <- model.matrix(object=Aformula,data=data) }
  if(method=="ML" | method=="DR" | method=="DRwt"){
    if(method=="DRwt")
```

```
      Yfit <- glm(formula=Yformula,family=Yfamily,data=data0,weights=w)
    else
      Yfit <- glm(formula=Yformula,family=Yfamily,data=data0)
    mu0 <- predict(object=Yfit,newdata=data0star,type="respons")
    eta0 <- predict(object=Yfit,newdata=data0star,type="link")
    nYpar <- length(Yfit$coef)
    LY <- model.matrix(object=Yformula,data=data) }

  #calculate estimate
  if(method=="ML") psi <- sum((Y-mu0)*A)/sum(A)
  if(method=="IPW") psi <- sum(Y*(A-(1-A)*w))/sum(A)
  if(method=="DR" | method=="DRwt") psi <- sum((Y-mu0)*(A-(1-A)*w))/sum(A)

  #calculate standard error
  if(method=="ML" | method=="DR" | method=="DRwt"){
    Yres <- matrix(0,nrow=n,ncol=nYpar)
    #must include those with A==1 as well here
    Yres[A==0,] <- estfun(Yfit); g <- family(Yfit)$mu.eta
    dmu.deta <- g(eta0); deta.dbeta <- LY
    dmu.dbeta <- dmu.deta*deta.dbeta
}
  if(method=="IPW" | method=="DR"  | method=="DRwt"){
    Ares <- estfun(Afit)
  }
  if(method=="ML"){
    psires <- A*(Y-mu0-psi); res <- cbind(psires,Yres)
    psiI <- c(sum(-A),colSums(-A*dmu.dbeta))/ncluster
    YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
      -solve(vcov(object=Yfit))/ncluster)
    I <- rbind(psiI,YI)
  }
  if(method=="IPW"){
    psires <- A*(Y-psi)-(1-A)*w*Y; res <- cbind(psires,Ares)
    psiI <- c(sum(-A),colSums(-(1-A)*Y*w*LA))/ncluster
    AI <- cbind(matrix(rep(0,nApar),nrow=nApar,ncol=1),
      -solve(vcov(object=Afit))/ncluster)
    I <- rbind(psiI,AI)
  }
  if(method=="DR"  | method=="DRwt"){
    psires <- A*(Y-mu0-psi)-(1-A)*w*(Y-mu0)
    res <- cbind(psires,Yres,Ares)
    psiI <- c(sum(-A),
      colSums(((1-A)*w-A)*dmu.dbeta),
      colSums(-(1-A)*(Y-mu0)*LA*w))/ncluster
    if(method=="DR")
      YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
        -solve(vcov(object=Yfit))/ncluster,
        matrix(rep(0,nYpar*nApar),nrow=nYpar,ncol=nApar))
    if(method=="DRwt")
      YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
        -solve(vcov(object=Yfit))/ncluster,
        t(Yres)%*%LA/ncluster)
    AI <- cbind(matrix(rep(0,nApar),nrow=nApar,ncol=1),
      matrix(rep(0,nApar*nYpar),nrow=nApar,ncol=nYpar),
      -solve(vcov(object=Afit))/ncluster)
    I <- rbind(psiI,YI,AI)
  }
  if(!missing(cluster))
    res <- aggregate(x=res,by=list(data[,cluster]),FUN=sum)[,-1]
  J <- var(res)
  se <- sqrt((solve(I)%*%J%*%t(solve(I))/ncluster)[1,1])

  #output
  out <- list(psi=psi,se=se); return(out)

}
```

References

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267.

Abadie, A. & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76**, 1537–1557.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources* **8**, 436–455.

Chen, X., Hong, H. & Tarozzi, A. (2008). Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. Tech. rep., Cowles Foundation Discussion Paper.

Dehejia, R. H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* **64**, 605–654.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies* **65**, 261–294.

Heckman, J. J. & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics* **30**, 239–267.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2**, 259–278.

Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86**, 4–29.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *The American Economic Review* **101**, 532–537.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* **76**, 604–620.

Lei, Q. (2011). *Improved double-robust estimation of missing data and causal inference models and efficient estimation of the average treatment effect on the treated*. Ph.D. thesis, Harvard University.

Neugebauer, R. & van der Laan, M. J. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference* **137**, 419–434.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review* **14**, 693–709.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics* **2**, 1–26.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics*. New York: Springer, pp. 331–457.

Zhang, Z., Chen, Z., Troendle, J. F. & Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics* **68**, 697–706.

[*Received April* $20xx$. *Revised September* $20xx$]