The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 17

Estimation Based on Case-Control Designs with Known Prevalence Probability

Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:
van der Laan, Mark J. (2008) "Estimation Based on Case-Control Designs with Known Prevalence Probability," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 17.
DOI: 10.2202/1557-4679.1114

Estimation Based on Case-Control Designs with Known Prevalence Probability

Mark J. van der Laan

Abstract

Regular case-control sampling is an extremely common design used to generate data to estimate effects of exposures or treatments on a binary outcome of interest when the proportion of cases (i.e., binary outcome equal to 1) in the population of interest is low. Case-control sampling represents a biased sample of a target population of interest by sampling a disproportional number of cases. Case-control studies are also commonly employed to estimate the effects of genetic markers or biomarkers on binary phenotypes.

In this article we present a general method of estimation relying on knowing the prevalence probability, conditional on the matching variable if matching is used.

Our general proposed methodology, involving a simple weighting scheme of cases and controls, maps any estimation method for a parameter developed for prospective sampling from the population of interest into an estimation method based on case-control sampling from this population.

We show that this case-control weighting of an efficient estimator for a prospective sample from the target population of interest maps into an efficient estimator for matched and unmatched case-control sampling. In particular, we show how application of this generic methodology provides us with double robust locally efficient targeted maximum likelihood estimators of the causal relative risk and causal odds ratio for regular case control sampling and matched case control sampling.

Various extensions and generalizations of our methods are discussed.

KEYWORDS: case control sampling, canonical gradient, causal effect, counterfactual, double robust estimation, efficient influence curve, estimating function, gradient, incidence density sampling, influence curve, inverse probability of treatment weighting, locally efficient estimation, marginal structural models, matched case control sampling, randomization assumption, randomized trial, semi-parametric regression, targeted maximum likelihood estimation

Author Notes: We are thankful for the helpful comments of the reviewers.

1 Introduction.

Case-control sampling is an extremely common design used to generate data to estimate effects of exposures or treatments on a binary outcome of interest when the actual population proportion of cases (i.e. binary outcome equal to 1) is small. As a consequence, it is of interest to present estimators of causal effects or variable importance parameters based on case-control data.

1.1 Formulation of case-control estimation problem.

Let's first formulate the statistical problem. For the sake of concreteness and illustration, our formulation will focus on a case-control point treatment data structure with baseline covariates in which one is concerned with estimation of the causal effect or variable importance of the treatment variable on the binary outcome. Our initial formulation will assume that the variables are not subject to missingness or censoring. Our general methods are straightforward extensions and apply to general case control data structures, including censored data structures and time-dependent longitudinal data structures.

Experimental unit of interest. Let $O^* = (W, A, Y) \sim P_0^*$ represent the experimental unit and corresponding distribution P_0^* of interest, consisting of baseline covariates W, a subsequent monitored treatment/exposure variable A, and a "final" binary outcome Y.

Causal or variable importance parameter of interest. Suppose one is concerned with statistical inference regarding a particular euclidean valued variable importance or causal effect parameter $\psi_0^* = \Psi^*(P_0^*) \in \mathbb{R}^d$ of this distribution P_0^*. For example, one might be interested in the marginal causal additive effect of a binary treatment $A \in \{0, 1\}$ defined as

$$\begin{aligned} {}^*_0 &\equiv E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\} = E_0^*(Y_1) - E_0^*(Y_0) \\ &= P_0^*(Y_1 = 1) - P_0^*(Y_0 = 1), \end{aligned}$$

where the latter causal effect interpretation of this parameter of P_0^* requires the notion of treatment specific counterfactual outcomes Y_0, Y_1, viewing $(W, A, Y = Y_A)$ as a time-ordered missing data structure on the full data structure (W, Y_0, Y_1), and one needs to assume the randomization assumption stating that A is independent of Y_0, Y_1, given W. The latter causal parameter formulation ${}^*_0$ can also be viewed as a W-adjusted variable importance (of variable A) parameter of the true regression of Y on A, W, in which case there is no need

to assume the time ordering ($W \Rightarrow A \Rightarrow Y$), the missing data structure assumption, or the randomization assumption, and the adjustment set W is user supplied (and does thus not need to correspond with the set of all confounders of A): see van der Laan (2006) for a general formulation of variable importance parameters and its direct relation to causal effect parameters.
One can also define the parameter of interest as a causal relative risk

$$\Psi_0^* = \frac{E_0^* E_0^*(Y \mid A=1, W)}{E_0^* E_0^*(Y \mid A=0, W)} = \frac{EY_1}{EY_0} = \frac{P(Y_1=1)}{P(Y_0=1)},$$

or a causal odds ratio,

$$\Psi_0^* = \frac{P(Y_1=1)P(Y_0=0)}{P(Y_1=0)P(Y_0=1)},$$

or their variable importance analogue.

We will use these particular marginal causal effects or marginal variable importance parameters as our main examples in order to illustrate our proposed methodology for case-control data, including our proposed targeted maximum likelihood estimation methodology.

Model for target probability distribution. A model for O^* is obtained by modelling this distribution of O^*: for example, one might know that A is independent of W, one might know the actual distribution (treatment mechanism) $P_0^*(A=a \mid W)$, or one might assume a marginal structural model

$$E_0^*(Y_a \mid V) = E_0^*(E_0^*(Y \mid A=a, W) \mid V) = m(a, V \mid \beta_0^*),$$

where $V \subset W$ denotes some user supplied potential effect modifier of interest, and $m(\cdot \mid \beta)$ some parameterization modelling the causal effect of the intervention $A=a$ on the outcome Y, conditional on V. If one wishes to avoid making causal assumptions, the marginal structural parameter represents the effect of a change in variable A on the mean outcome of Y within subgroups $V=v$, controlling for potential confounders W. We will denote such a model for P_0^* with $\mathcal{M}^*$: i.e., it is assumed that $P_0^* \in \mathcal{M}^*$.

Case-control sampling and its probability distribution. If one would sample n i.i.d. observations $O_1^*, \ldots, O_n^* \sim P_0^*$, then we could (e.g.) apply the locally efficient targeted MLE of ψ_0^* (see e.g. van der Laan and Rubin (2006) or Moore and van der Laan (2007)), or one could use double robust estimating function methodology (van der Laan and Robins (2002)).

DOI: 10.2202/1557-4679.1114

However, this so called prospective sampling scheme is often considered impractical and ineffective in situations in which the probability $P_0^*(Y=1)$ on the event $Y=1$ (say disease) is very small. For example, if the proportion of diseased in the population of interest is one in hundred thousand, then one would have to sample millions of observations in order to have some cases (i.e, $Y_i=1$) in the sample. This sparsity of cases in the population of interest is precisely the typical motivation for case-control sampling.

We will distinguish between two types of case-control sampling: independent or un-matched case-control sampling and matched case-control sampling. In both cases, the marginal distribution of the cases and the marginal distribution of the controls is completely determined by the population (i.e. prospective sampling) distribution P_0^* of the random variable (W,A,Y) of interest.

Independent Case-Control Sampling. One first samples a *case* by sampling (W_1,A_1) from the conditional distribution of (W,A), given $Y=1$. Subsequently, one samples J *controls* (W_0^j,A_0^j) from the conditional distribution of (W,A), given $Y=0$, $j=1,\ldots,J$. It is allowed that these J control observations are dependent as long as their marginal distributions are indeed equal to the conditional distribution of W,A, given $Y=0$.

This results in an experimental unit observed data structure:

$$O=((W_1,A_1),(W_0^j,A_0^j:j=1,\ldots,J))\sim P_0,$$

where we denote the sampling distribution of this data structure O described above with P_0. Thus, a case control data set will consists of n independent and identically distributed observations $O_1,\ldots,O_n$ with sampling distribution P_0 described above. That is, we treat the cluster consisting of one case and J controls as the experimental unit, and the marginal distribution of the case and controls are specified as above by P_0^*.

Matched Case-Control Sampling. One specifies a categorical matching variable $M\subset W$. One first samples a case by sampling (M_1,W_1,A_1) from the conditional distribution of (M,W,A), given $Y=1$. Subsequently, one samples J controls (M_0^j,W_0^j,A_0^j) from the conditional distribution of (M,W,A), given $Y=0,M=M_1$. That is, with probability equal to 1 we have $M_0^j=M_1$, $j=1,\ldots,J$. It is allowed that these J control observations are dependent as long as their marginal distributions are indeed equal to the conditional distribution of M,W,A, given $Y=0,M=M_1$.

This results in an experimental unit data structure:

$$O=((M_1,W_1,A_1),(M_0^j=M_1,W_0^j,A_0^j:j=1,\ldots,J))\sim P_0,$$

where we denote the sampling distribution of this data structure O described above with P_0. Thus, a matched case-control data set will consists of n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0 described above. That is, we treat the cluster consisting of one case and the J matched controls as the experimental unit, and the marginal distribution of the case and J controls are specified as above by P_0^*

We will also refer to the independent case-control experiment and the matched case-control experiments as Case-Control Design I and Case-Control Design II, respectively.

Extensions. Our methods naturally handle the case that J is random and thus varies per experimental unit, assuming that the marginal distributions of cases and controls, conditional on $J = j$, do not depend on j. In the situation that a case was never coupled to a set of controls one can artificially create such couplings, and apply our methods, and one could average over a variety of sensible coupling schemes. The latter shows that if the true independent case control design simply involves sampling a set of cases and an independent set of controls, without any coupling, then our case control weighting methods show that one should weight each case by q_0 and each control by $(1 - q_0)/\bar{J}$, where $\bar{J}$ is the number of controls divided by the number of cases. In the discussion we show the simple extension of our methods to some variations on these case-control designs I and II, such as pair-matched case-control designs, case-control sampling within strata, and counter-match case control designs. We also note here that our sampling model for O^* corresponds with sampling with replacement from a particular population with population distribution P_0^*. Such a model is appropriate if the size of the total population is large relative to sample size n.

The estimation problem. The statistical problem is now to estimate the parameter $\psi_0 = \Psi^*(P_0^*)$ of the population distribution $P_0^* \in \mathcal{M}^*$ of (W, A, Y), known to be an element of some specified model $\mathcal{M}^*$, based on the case-control data set $O_1, \ldots, O_n \sim P_0$.

Known or sensitivity analysis parameters/weights. We define

$$q_0 \equiv P_0^*(Y = 1) \text{ and } q_0(\delta \mid M) \equiv P_0^*(Y = \delta \mid M),$$

as the marginal probability of being a case, and the conditional probability of being a case/non-case, conditional on the matching variable. It is assumed that

these probabilities are between 0 and 1. In addition, we define the quantity

$$\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y=0 \mid M)}{P_0^*(Y=1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.$$

We note that $\bar{q}_0(M)$ is determined by q_0 and $q_0(1 \mid M) = P_0^*(Y=1 \mid M)$, and we also note that $E_0 \bar{q}_0(M_1) = 1 - q_0$. These two quantities q_0 and $\bar{q}_0(M)$ (for matched case-control studies) will be used to weight the cases and controls to obtain valid estimation procedures.

In order to be able to identify the wished causal parameters, for case-control design I, we only need to assume q_0 is known, and, for matched case-control design II, we assume q_0 and $\bar{q}_0(m)$ for each m are known. However, we note here that for matched case-control designs one can also assume that q_0 and

$$r_0(m) \equiv P_0^*(Y=0, M=m)$$

(instead of $\bar{q}_0(1 \mid m)$) are known We note that, given $r_0(m)$, $\bar{q}_0(m)$ is known up till a simple to estimate nuisance parameter $P(M_1 = m)$:

$$\bar{q}_0(m) = \frac{r_0(m)}{P_0(M_1 = m)}.$$

As a consequence, our case-control weighted estimation procedures using q_0, $\bar{q}_0(m)$ still apply in settings in which one assumes q_0 and $r_0(m)$ are known, by replacing $\bar{q}_0(m)$ by its estimate $\frac{r_0(m)}{\frac{1}{n}\sum_{i=1}^n I(M_{1i}=m)}$.

Observed data model. In this article, we will assume that q_0 is known, and that, for matched case-control designs we also assume that $\bar{q}_0(M)$, or equivalently, $q_0(1 \mid m) = P_0^*(Y=1 \mid M=m)$ is known for each m. In our accompanying technical report we show that if the "treatment mechanism" $g_0^*(a \mid w) = P_0^*(A=a \mid W=w)$ is known, as it would be in a case control study nested in a randomized trial, then we can estimate the relative risk or odds ratio parameters without a need to know (any of) q_0 or $\bar{q}_0(M)$.

The model $\mathcal{M}^*$, possibly including the knowledge q_0 or $\bar{q}_0(M)$, imply now models for the marginal distribution of the cases (M_1, W_1, A_1) and the marginal distributions of the controls (M_1, W_2^j, A_2^j), $j = 1, \ldots, J$. The model $\mathcal{M}^*$ does not imply much, if anything, about the dependence structure among (M_1, W_1, A_1), (M_1, W_2^j, A_2^j), $j = 1, \ldots, J$, beyond the fact that, for matched case-control studies, all its components (i.e., the case and control observations) share a common variable M_1. Let $\mathcal{M}$ be the model for the observed

data distribution P_0 compatible with $\mathcal{M}^*$ (i.e., its marginals are specified by P_0^*).

One possible and probably very common model $\mathcal{M}$ is to assume that, given the first draw (M_1, W_1, A_1) from (M, W, A), given $Y = 1$, the control observations are all *independent* draws from the specified conditional distributions. Note that in this latter model the marginal distributions for the case and control observations implied by P^* describe now the whole case-control sampling distribution P, so that we can write $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}\}$, where $P(P^*)$ is the distribution of O implied by P^*.

Other possible models might specify in another manner, or not specify at all, the dependence structure and could, for example, be represented as $\{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$, where the nuisance parameter η in combination with P^* describes the complete joint distribution of case and control observations $(M_1, Z_1), (M_1, Z_2^j : j = 1, \ldots, J)$ compatible with its marginal distributions implied by P^*.

We note that knowing q_0 does not put restrictions on the data generating distribution P_0 since one conditions on $Y = 1$, but for case-control design I it does allow identification of the wished parameters by expressing them as a function of the distribution of the observed case-control data-structure and q_0. Similarly, for matched case-control designs, knowing q_0 and $r_0(\cdot)$ does not put restrictions on the data generating distribution P_0 for matched case-control designs, but it allows one to express the wished parameter as a function of the distribution of the data and (q_0, r_0). It remains to be investigated if knowing q_0 and $\bar{q}_0$ puts a restriction on the data generating distribution for matched-case-control designs.

1.2 Overview of article.

In Section 2 we present our general solution to the estimation problem for these two types of case control designs I and II, which weights the cases and controls with q_0 and $(1 - q_0)/J$ ($\bar{q}_0(M)/J$ for case control design II), respectively, and then applies a method developed for prospective sampling to estimate the parameter of interest (e.g., targeted maximum likelihood estimators or estimating equations for the causal effect or variable importance parameter ψ_0 of interest), as if the data was directly drawn from the population distribution P_0^* of interest. In other words, each estimating function for ψ_0^* or likelihood for P_0^* in the underlying model $\mathcal{M}^*$ maps into a "case-control"-weighted estimating function or likelihood for the observed data model $\mathcal{M}$ (whatever nuisance parameter specification $P(P^*, \eta)$ it might have beyond the description of its marginal distributions in terms of P^*).

DOI: 10.2202/1557-4679.1114

Beyond the weighting, we point out that one should aim to select the best among these case-control weighted estimating equations/procedures for the observed case-control data. In Section 3 we show the important and convenient result that case-control weighting of the efficient procedure for the parameter of interest (as formalized by the efficient influence curve) in the prospective sampling model $\mathcal{M}^*$ maps into the efficient procedure for the observed case-control data model $\mathcal{M}$. This implies, in particular, that case-control weighting of the locally efficient targeted maximum likelihood estimator developed for prospective sampling model $\mathcal{M}^*$ results in a locally efficient targeted maximum likelihood estimation procedure for case-control sampling. In general, the power of our generic method is that one can map the estimation procedures developed for prospective sampling into highly or fully efficient estimation procedures for case-control sampling. In particular, our method is now able to fully exploit software developed for prospective sampling.

To summarize, in Section 2 and Section 3 we establish general properties of our case-control weighted mapping from estimating functions/influence curves/gradients for the parameter of interest for model $\mathcal{M}^*$ into estimating functions/influence curves/gradients for the parameter of interest for the observed data model $\mathcal{M}$, showing that 1) the case-control weighting does map each parameter-specific influence curve for the model $\mathcal{M}^*$ into a parameter-specific influence curve for model $\mathcal{M}$, 2) it maps the efficient influence curve/ canonical gradient for model $\mathcal{M}^*$ into the efficient influence curve/canonical gradient for model $\mathcal{M}$, and 3) that our case-control weighting inherits any robustness of estimating functions/influence curves for model $\mathcal{M}^*$.

We suggest that even in cases that q_0 (or $q_0(1 \mid M)$ for matched case control designs) is unknown, it is of interest to present these estimators and inferences for an interval of possible q_0-values, thereby presenting a sensitivity analysis.

As an example we show that indeed for case-control design I the case-control weighted targeted maximum likelihood estimator is indeed a locally efficient double robust estimator. This implementation of a targeted maximum likelihood estimators needs to guarantee that the initial maximum likelihood fit of the logistic regression $P_0^*(Y = 1 \mid A, W)$ is proportional to q_0, which is a requirement for these double robust estimators to *not* suffer from a large variance due to the singularity $q_0 \approx 0$. The latter is precisely guaranteed by our case-control weighting method.

These double robust targeted maximum likelihood estimators rely on knowing the incidence probability q_0 and, for case-control design II, $\bar{q}_0(M)$, beyond either a correctly specified model for $Q^*(A, W) = P_0^*(Y = 1 \mid A, W)$ or a correctly specified model for $g_0^*(a \mid W) = P_0^*(A = a \mid W)$.

In Section 4, we end this article with a discussion and point out a number

of extensions. Various technical proofs are deferred to the Appendix.

1.3 Some relevant literature.

Case-control studies are probably one of the most commonly used designs, if not the most used design. For example, searching for case-control analysis on PubMed resulted in a list of 56,000 articles. Their use is not limited to public health applications; case-control studies are also frequently performed in econometric applications (See Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981)). Logistic regression is the most commonly used model in the literature for case-control studies. Conditional logistic regression is the prominent method in the literature for matched case-control studies and the statistical methodology goes back to the early 80's.

We will discuss these two methods briefly as well as related IPTW methods, as it goes without saying that an overview of the literature in this area is not possible. However, our proposed general methodology is not covered by the current literature, as far as we know.

Some of the key papers on logistic regression in standard case-control studies are Anderson (1972), Prentice and Pyke (1979), Breslow (1996), and Breslow and Day (1980). Breslow et al. (2000) establish asymptotic efficiency of the standard maximum likelihood estimator ignoring the case-control sampling. The most frequently cited sources for conditional logistic regression for matched case-control studies are Breslow and Day (1980), Holford et al. (1978), and Breslow et al. (1978). Various books considering case-control studies are Schlesselman (1982), Collett (1991), Jewell (2004), Rothman and Greenland (1998), and Hosmer and Lemeshow (2000), among others.

Cohort studies differ from case-control studies in that they sample exposed ($A = 1$) and unexposed ($A = 0$) individuals rather than diseased ($Y = 1$) and non-diseased ($Y = 0$). When cohort studies are matched, they are matched based on the exposure variable in an effort to reduce the bias found in observational studies. There has been much work in this area, particularly in the analysis and matching of cohort studies, by W.G. Cochran, D.B. Rubin, P.R. Rosenbaum, and N. Thomas. A collection of this work can be found in Rubin (2006). A thorough discussion of cohort study design can also be found in Rothman and Greenland (1998).

The method of adding an intercept to a standard logistic regression fit, and, in that manner, estimating effects different from the odds-ratio has been presented in the literature (see e.g. Anderson (1972), Prentice and Breslow (1978), Greenland (1981), Morise et al. (1996), Wachholder (1996), Greenland (2004)).

DOI: 10.2202/1557-4679.1114

Matched case-control studies are most frequently handled with conditional logistic regression models, but these designs and methods also have limitations. Firstly, it does not allow estimation of the effect of the matching variable on the diease (see, Schlesselman (1982), Rothman and Greenland (1998)): Any variable used for matching cannot be studied as a risk factor, since cases and controls are constrained to be equal with respect to the variables that are matched. Secondly, matching can hurt the precision if the matching variable is correlated with the exposure variable and not disease, which is often called over-matching. Finally, as we remarked from the start, these methods are by necessity heavily model based, while the methods presented here, relying on knowing the case-control weights, allow double robust locally efficient estimation in semiparametric models, thereby allowing the use of methods which minimize the reliance of the inference on unknown model assumptions.

Robins (1999) discusses the approximately correct IPTW-method for estimation of the unknown parameters in a marginal structural logistic regression model for a direct effect analysis based on standard case-control data under the assumption that the population proportion of cases, q_0, is small. We also refer to Newman (2006) for an IPTW-type approach for fitting marginal structural models based on case-control data. Mansson et al. (2007) investigate a variety of IPTW and propensity score methods in case-control studies through a simulation study, which includes the IPTW estimator for the logistic marginal structural model.

Notation. We introduce now some useful notation. Let $O^* \rightarrow D^*(O^*)$ represent an estimating function or loss function for O^* that can thus be used to estimate the parameter of interest of P_0^* based on an i.i.d sample of O^*. This article is concerned with mapping this function D^* into an estimating function of loss function for this same parameter of interest, but now based on sampling O (i.e., a biased sample for O^*). Given such a function $D^*(O^*)$, we define a case-control weighted version $D_{q_0}(O) \equiv q_0 D^*(W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1) D^*(W_2^j, A_2^j, 0)$ of D^*, which is now a function of the observed experimental unit O. We define the expectation operator $P_{0,q_0} D^* = P_0 D_{q_0}$, which thus simply takes the expectation of the case-control weighted function $D_{q_0}(O)$ w.r.t. P_0. Similarly, we define the empirical expectation $P_{n,q_0} D^* = P_n D_{q_0}$ as the empirical mean of the case-control weighted D_{q_0}, where P_n is the empirical distribution of $O_1, \ldots, O_n$. We apply this notation to both case-control designs, where for case-control design I $\bar{q}_0(M_1)$ reduces to $1 - q_0$.

2 Case-Control weighting of estimation procedures developed for prospective sampling.

Throughout this section, we will make the convention that $\bar{q}_0(M)$ reduces to $1 - q_0$ in the case control design I, so that we can state our results for both the regular case-control design I and the matched case-control design II in one formula.

We start out with stating the theorem which proves that the case-control weighting maps a function of O^* into a function of the case-control data structure O, while preserving the expectation of the function.

Definition 1 *(***Case-control weighted function***) Given a* $D^*(O^*) = D^*(W, A, Y)$ *we define the case-control weighted version of* D^* *as*

$$D_{q_0}(O) \equiv q_0 D^*(M_1, W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1) D^*(M_1, W_2^j, A_2^j, 0),$$

where in the special case of Case Control Design I, we have $\bar{q}_0(M) = 1 - q_0$.

Theorem 1 *(***Unbiased estimating function mapping***) Let* $D^*(O^*) = D^*(W, A, Y)$ *be a function so that* $P_0^* D^* \equiv E_{P_0^*} D^*(O^*) = 0$. *Then* $P_0 D_{q_0} = 0$. *In particular, in Case Control Design I,*

$$D_{q_0}(0) \equiv q_0 D^*(W_1, A_1, 1) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} D^*(W_2^j, A_2^j, 0)$$

satisfies $P_0 D_{q_0} = 0$.

In more generality, for any function D^* *and corresponding case control weighted function* D_{q_0}, *we have*

$$P_0 D_{q_0} = P_0^* D^*.$$

Proof. We provide the proof for case-control design II and we suppress the index q_0 in D_{q_0}. The same proof applies to case-control design I. First, we note that $P_0 q_0 D(M_1, W_1, A_1, 1) = \int_{M_1, W_1, A_1} D(M_1, W_1, A_1, 1) P_0^*(M_1, W_1, A_1, Y = 1)$. Secondly, we note that

$$P_0 \bar{q}_0(M_1) D(M_1, W_2^j, A_2^j, 0) =$$
$$\int_{m,w,a} D(m, w, a, 0) \bar{q}_0(m) P_0(M_1 = m) P_0^*(W = w, A = a \mid M = m, Y = 0),$$

where we also need to note that $P_0(M_1 = m) = P_0^*(M = m \mid Y = 1)$. We have

$$\begin{aligned}
&\bar{q}_0(m) P_0(M_1 = m) P_0^*(W = w, A = a \mid M = m, Y = 0) \\
&= \frac{\bar{q}_0(m) P_0^*(M=m|Y=1) P_0^*(W=w, A=a, M=m, Y=0)}{P_0^*(Y=0, M=m)} \\
&= P_0^*(M = m, W = w, A = a, Y = 0).
\end{aligned}$$

This proves that

$$\begin{aligned}
&P_0 D = \int_{M_1,W_1,A_1} D(M_1, W_1, A_1, 1) P_0^*(M_1, W_1, A_1, Y = 1) \\
&+ \frac{1}{J} \sum_{j=1}^{J} \int_{M_1,W_2,A_2} D(M_1, W_2, A_2, 0) P_0^*(M_1, W_2, A_2, Y = 0) \\
&= P_0^* D = 0.
\end{aligned}$$

This completes the proof. □

In the next section we establish general properties of this mapping which help us to understand the generality and optimality of the statistical approach for dealing with case-control sampling implied by this mapping. In this section we focus on the statistical (i.e., methodological) implications of this mapping for the analysis of case-control data,

2.1 Preservation of robustness of case-control weighted functions.

If a function D^* satisfying $P_0^* D(P_0^*) = 0$ also satisfies the robustness property $P_0^*(D(P^*)) = 0$ for any $P^* \in \mathcal{M}_1^* \subset \mathcal{M}^*$ for a submodel $\mathcal{M}_1^*$, then the same robustness w.r.t. to misspecification of P_0^* applies to D_{q_0} since, for $P^* \in \mathcal{M}_1^*$, $P_0 D_{q_0}(P^*) = P_0^* D(P^*) = 0$.

In particular, double robust estimating functions for censored and causal inference data structures and models $\mathcal{M}^*$, as presented in general in van der Laan and Robins (2002), are mapped into double robust case-control weighted estimating functions.

In the remainder of this section we outline the general statistical methods implied by the case-control weighted mapping. Estimating function methodology developed for prospective sampling immediately implies now, through the case-control weighted mapping, estimating function methodology for case-control sampling. In particular, in view of the general estimating function theory presented in van der Laan and Robins (2002) it follows that the case control mapping is a mapping from estimating functions (or gradients, see van der Laan and Robins (2002)) developed for a model for P_0^* into estimating functions based on case-control sampling from P_0. For details we refer to our technical report, and here we suffice with an illustration.

2.2 Example: Case-control weighted double robust estimating function.

Let's illustrate this estimating function method by constructing a double robust estimator of the additive causal effect $\psi_0^* = E(Y_1 - Y_0)$ for a nonparametric model $\mathcal{M}^*$ for the distribution P_0^* of (W, A, Y). Let $g_0^*(A \mid M, W)$ denote the conditional distribution of A, given W, and let $Q_0^*(M, W, A)$ denote the conditional probability of Y, given M, W, A, under P_0^*.

The double robust efficient estimating function for sampling from P_0^* is given by

$$\begin{aligned} D^*(\psi^*, g^*, Q^*)(O^*) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} - \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} (Y - Q^*(M, W, A)) \\ &\quad + Q^*(M, W, 1) - Q^*(M, W, 0) - \psi^*, \end{aligned} \tag{1}$$

where g^* and Q^* represent candidates for the nuisance parameters g_0^* and Q_0^* of this estimating function for ψ_0^*.

It is double robust in the sense that

$$E_0^* D^*(\psi_0^*, g^*, Q^*)(O^*) = 0 \text{ if either } g^* = g_0^* \text{ or } Q^* = Q_0^*,$$

and in both cases one needs that $g^*(1 \mid W) g^*(0 \mid W) > 0$ a.e. Let $D^*(g^*, Q^*)$ be defined so that $D^*(\psi^*, g^*, Q^*) = D^*(g^*, Q^*) - \psi^*$.

The weighted double robust estimating function for case-control data is thus given by:

$$\begin{aligned} D_{q_0}(\psi^*, g^*, Q^*)(O) &= q_0 D^*(\psi^*, g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(\psi^*, g^*, Q^*)(M_1, W_2^j, A_2^j, 0), \end{aligned}$$

or we can define it as

$$\begin{aligned} D_{q_0}(\psi^*, g^*, Q^*)(O) &= q_0 D^*(g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(g^*, Q^*)(M_1, W_2^j, A_2^j, 0) - \psi^*. \end{aligned}$$

This estimating function is now also double robust for case control data:

$$E_0 D_{q_0}(\psi_0^*, g^*, Q^*) = 0 \text{ if either } g^* = g_0^* \text{ or } Q^* = Q_0^*,$$

and in both cases one needs that $g^*(1 \mid W) g^*(0 \mid W) > 0$ a.e.

DOI: 10.2202/1557-4679.1114

The solution ψ_n of the case-control weighted estimating equation:

$$P_n D_{q_0}(g_n^*, Q_n^*) - \psi^* = 0$$

exists in closed form and is given by:

$$\begin{aligned} {}_n &= \frac{1}{n}\sum_{i=1}^{n} q_0 D^*(g_n^*, Q_n^*)(M_{1i}, W_{1i}, A_{1i}, 1) \\ &\quad + \frac{\bar{q}_0(M_{1i})}{J}\sum_{j=1}^{J} D^*(g_n^*, Q_n^*)(M_{1i}, W_{2i}^j, A_{2i}^j, 0). \end{aligned}$$

This estimator is now consistent if either g_n^* consistently estimates g_0^* or Q_n^* consistently estimates Q_0^*, which explains why it is called double robust.

Under some extra appropriate regularity conditions, this estimator is also asymptotically linear and thereby has a normal limit distribution (see van der Laan and Robins (2002) for general "central limit" theorems for solutions of estimating equations). In particular, if g_n^* consistently estimates g_0^* and Q_n^* consistently estimates Q_0^*, then, under appropriate regularity conditions, ${}_n$ is asymptotically linear with influence curve $D_{q_0}(g_0^*, Q_0^*, \psi_0)$ and is thus asymptotically efficient. The estimators g_n^* and Q_n^* can be based on case-control weighting of maximum likelihood estimators for the prospective model, as presented in next subsection.

Statistical behavior of double robust estimator when cases are rare. Inspection of this influence curve D_{q_0} sheds some light on the statistical behavior of this double robust estimator for the important case that $q_0 \approx 0$ is very small. In particular, we are interested in how well one can estimate the relative effect ψ_0/q_0, since ψ_0 is itself very small. It follows that, in general, the influence curve of ψ_n/q_0 as an estimator of ψ_0/q_0 will blow up for small values q_0, *except if it guaranteed that* $Q_n^* = q_0 Q_n^{\#}$ *for some bounded estimator* $Q_n^{\#}$. Therefore, in our proposed targeted maximum likelihood or double robust estimator we propose such estimators based on either case-control weighted logistic regression fits or intercept adjusted logistic regression fits (see Section 2 accompanying technical report).

2.3 Case-control weighted loss functions.

Our case-control weighting can also be used to map loss functions for the underlying model $\mathcal{M}^*$ into loss functions for the observed data model $\mathcal{M}$. In particular, we can construct a case-control weighted log likelihood loss function.

Theorem 2 *(***Case Control Weighted Log-Likelihood Loss function***)* *Define the following case-control weighted log-likelihood loss function for the density p_0^* of O^* under sampling of $O \sim P_0$:*

$$L(p^*, O) = q_0 \log p^*(M_1, Z_1, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^{J} \log p^*(M_1, Z_2^j, 0).$$

In particular, in Case Control Design I, we have

$$L(p^*, O) = q_0 \log p^*(M_1, Z_1, 1) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \log p^*(M_1, Z_2^j, 0).$$

We have

$$p_0^* = \arg\max_{p^*} E_0 L(p^*, O),$$

where the argmax is taken over all densities p^*. That is, the density maximizing the expectation of the loss function $L(p^*, O)$ is unique and given by the density p_0^* of O^*.

The proof of this theorem is similar to the proof of Theorem 1 and is therefore omitted.

2.4 Case-control weighted maximum likelihood estimation.

Given a specified model $\mathcal{M}^*$ for p_0^*, we can estimate P_0^* with the case-control weighted maximum likelihood estimator:

$$p_n^* = \arg\max_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} L(O_i, p^*).$$

The implementation of this weighted maximum likelihood estimator simply involves assigning weights q_0 to the cases, assigning weights $\bar{q}_0(M_{1i})/J$ to the corresponding J controls, and then implementing the maximum likelihood estimator for prospective sampling (i.e. treating the sample of cases and controls as an i.i.d sample of P_0^*), thus ignoring the case control sampling.

For example, let's consider the point treatment data structure $O^* = (M, W, A, Y)$. Consider a nonparametric model for the marginal distribution of W, Q_W^*, a model $\{g_\eta^* : \eta\}$ for $g_0^*(A \mid M, W)$, and a model $\{Q_\theta^* : \theta\}$ for the conditional distribution $P_0^*(Y = 1 \mid M, W, A) = Q_0^*(M, W, A)$.

The case-control weighted maximum likelihood estimator of the marginal distribution of W is now the weighted empirical distribution of the pooled sample $(W_{1i}, (W_{2i}^j : j = 1, \ldots, J))$. Similarly, the case-control weighted maximum likelihood estimator of $g_0^*(A \mid W)$ is given by

$$\eta_n = \arg\max_{\eta} \sum_{i=1}^{n} q_0 \log g_\eta^*(A_{1i} \mid M_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log g_\eta^*(A_{2i}^j \mid M_{1i}, W_{2i}^j),$$

and the case-control weighted maximum likelihood estimator of $Q_0^*(M, W, A)$ is given by

$$\theta_n = \arg\max_{\theta} \sum_{i=1}^{n} q_0 \log Q(M_{1i}, W_{1i}, A_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q(M_{1i}, W_{2i}^j, A_{2i}^j)).$$

Indeed, it follows that each of these case-control weighted maximum likelihood estimators can be implemented by assigning the two weights q_0 and $\bar{q}_0(M_1)$ to the cases and controls, respectively, and apply the standard maximum likelihood estimator of the density p_0^* under prospective sampling.

Given the weighted maximum likelihood estimators Q_{1n}^* and Q_n^*, described above, the corresponding substitution estimator of $EY_a = E_{Q_1^*} Q^*(W, a)$ is given by

$$_n(a) = \frac{1}{\sum_{i=1}^{n} \{q_0 + \bar{q}_0(M_{1i})\}} \sum_{i=1}^{n} q_0 Q_n^*(M_{1i}, W_{1i}, a) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} Q_n^*(M_{1i}, W_{2i}^j, a).$$

In particular, these estimators of EY_0 and EY_1 now map into an estimator $_n(1)/\psi_n(0)$ of the relative risk EY_1/EY_0.

2.5 Case-control weighted targeted maximum likelihood estimation.

Targeted maximum likelihood estimation is a general methodology introduced in van der Laan and Rubin (2006) and illustrated with a variety of examples. The case-control weighting allows us now to provide a case-control weighted targeted maximum likelihood estimation methodology targeting the parameter of interest.

Specifically, let $D^*(P_0^*)$ be the efficient influence curve of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$. Consider an initial estimator P_n^{*0} of P_0^* based on $O_1, \ldots, O_n$ such as a case-control weighted maximum likelihood estimator according to a working model within $\mathcal{M}^*$. Let $\{P_n^*(\epsilon) : \epsilon\}$ be a submodel of $\mathcal{M}^*$ with

parameter ϵ satisfying that the linear span of its score at $\epsilon = 0$ includes $D^*(P_n^{*0})$. Let ϵ_n^1 be the case-control weighted maximum likelihood estimator of ϵ:

$$\epsilon_n^1 = \arg\max P_{n,q_0} \log p_n^{*0}(\epsilon).$$

This yields an update $P_n^{*1} = P_n^{*0}(\epsilon_n^1)$ of the initial estimator P_n^{*0}. We iterate this updating process till step k at which $\epsilon_n^k \approx 0$ and we denote the final update with P_n^*. By the score condition, this final estimator solves the case-control weighted efficient influence curve:

$$0 = P_{n,q_0} D^*(P_n^*) = P_n D_{q_0}(P_n^*)$$

up till numerical precision (see van der Laan and Rubin (2006)). We refer to $\psi_n = \Psi^*(P_n^*)$ as the case-control weighted targeted maximum likelihood estimator of ψ_0.

One particular approach for establishing the asymptotics of this estimator is obtained under the assumption that $D^*(P^*) = D^*(\psi^*, \eta^*)$ for some nuisance parameter, thereby assuming an estimating function representation for the efficient influence curve. (This assumption is not necessary at all to establish the same asymptotics: see van der Laan and Rubin (2006).) In this case, it follows that the targeted maximum likelihood estimator ψ_n solves $P_n D_{q_0}(\psi_n, \eta_n^*) = 0$ so that one can establish asymptotic linearity of ψ_n and derive its influence curve under relatively standard differentiability and empirical process conditions.

In particular, if η_n^* is a consistent estimator of a η_0^* satisfying $P_0 D_{q_0}(\psi_0, \eta_0^*) = 0$, then under such standard conditions, asymptotic consistency and asymptotic linearity can be established. For example, if $\eta_0^* = \eta(P_0^*)$ is the true parameter, then ψ_n will have influence curve given by $D_{q_0}(\psi_0, \eta_0^*)$.

2.6 Case-control weighted targeted MLE of marginal causal effect for case control data.

We will illustrate the targeted maximum likelihood estimator for the parameter $_0 = EY_1 - EY_0$ and the nonparametric model $\mathcal{M}^*$ for the point treatment data structure $(W, A, Y) \sim P_0^*$.

Recall that the double robust estimating function/efficient influence curve of Ψ under i.i.d sampling from P_0^* is given by

$$\begin{aligned} D^*(g^*, Q^*)(M, W, A, Y) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} \\ &\quad \times (Y - Q_2^*(M, W, A)) \end{aligned}$$

$$\begin{aligned} & +Q_2^*(M,W,1) - Q_2^*(M,W,0) - \Psi(Q^*) \\ \equiv \quad & D_1^*(g^*,Q^*)(M,W,A,Y) + D_2^*(Q^*)(M,W), \end{aligned}$$

where $Q^* = (Q_1^*, Q_2^*)$ represents both the marginal distribution Q_1^* of W and the conditional distribution Q_2^* of Y, given A, W. We note that $D^*(g^*, Q^*)$ can also be represented as an estimating function for ψ since $D^*(g^*, Q\ ) = D^*(\Psi(Q^*), g^*, Q^*)$, as we did above.

Let Q_{2n}^{*0} be an initial estimator of $Q_{20}^*(A,W) = P_0^*(Y=1 \mid A,W)$ according to a particular working model $\mathcal{Q}^w$ for Q_{20}^*: for example,

$$Q_{2n}^{*0} = \arg \max_{Q_2^* \in \mathcal{Q}^w} \sum_{i=1}^n q_0 \log Q_2^*(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log(1 - Q_2^*(A_{2i}^j, W_{2i}^j)),$$

or the logistic regression based estimator Q_{n,q_0}^* using an intercept adjustment in terms of $\log q_0/(1-q_0)$ presented in Section 2 of the accompanying technical report.

Given a model $\mathcal{G}$ for g_0^*, let g_n^* be the corresponding weighted MLE:

$$g_n^* = \arg \max_{g \in \mathcal{G}} \sum_{i=1}^n q_0 \log g(A_{1i} \mid W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log g(A_{2i}^j \mid W_{2i}^j).$$

Similarly, let Q_{1n}^* be the nonparametric weighted MLE:

$$Q_{1n}^* = \arg \max_{Q_1} \sum_{i=1}^n q_0 \log dQ_1(W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log dQ_1(W_{2i}^j),$$

where the maximum is over all discrete distributions which put mass on W_{1i} and W_{2i}, $i = 1, \ldots, n$. It follows that Q_{1n}^* is a discrete distribution which puts mass q_0/n on W_{1i}, $i = 1, \ldots, n$, and puts mass $\bar{q}_0(M_{1i}))/(nJ)$ on W_{2i}^j, $j = 1, \ldots, j$, $i = 1, \ldots, n$.

Given any Q^*, g^*, let $\{Q_{2g^*}^*(\epsilon) : \epsilon\}$ be a model through Q_2^* at $\epsilon = 0$ and satisfying that the span of its score at $\epsilon = 0$ includes the component $D_1^*(g^*, Q^*)$ of the efficient influence curve of Ψ under i.i.d. sampling from P_{Q^*,g^*}^*. For example,

$$\frac{d}{d\epsilon} \log \left\{ Q_{2g^*}^*(\epsilon)^Y (1 - Q_{2g^*}^*(\epsilon))^{1-Y} \right\} \Big|_{\epsilon=0} = D_1^*(g^*, Q^*).$$

This can be achieved with the following fluctuation function of Q_2^*:

$$\text{logit} Q_{2g^*}^*(\epsilon) = \text{logit} Q_2^* + \epsilon Z(g^*),$$

where

$$Z(g^*) \equiv \left\{ \frac{I(A=1)}{g^*(1 \mid M,W)} - \frac{I(A=0)}{g^*(0 \mid M,W)} \right\}.$$

Given the estimator g_n^* of g_0^*, consider the fluctuation function $\{Q_{2ng_n^*}^{*0}(\epsilon) : \epsilon\}$ and let ϵ_n^0 be its weighted MLE:

$$\epsilon_n^0 = \arg\max_{\epsilon} \sum_{i=1}^{n} q_0 \log Q_{2ng_n^*}^{*0}(\epsilon)(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q_{2ng_n^*}^{*0}(\epsilon)(A_{2i}^j, W_{2i}^j)),$$

which can be computed with standard logistic regression software.

The first step targeted MLE is now defined as

$$(g_n^*, Q_{1n}^*, Q_{2n}^{*1} = (g_n^*, Q_{1n}^*, Q_{2n}^0(\epsilon_n^0)).$$

The k-th step targeted MLE is given by $(g_n^*, Q_{1n}^*, Q_{2n}^{*k} = Q_{2n}^{*k-1}(\epsilon_n^{k-1}))$, where, for $k = 0, \ldots$

$$\epsilon_n^k = \arg\max_{\epsilon} \sum_{i=1}^{n} q_0 \log Q_{2ng_n^*}^{*k}(\epsilon)(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q_{2ng_n^*}^{*k}(\epsilon)(A_{2i}^j, W_{2i}^j)).$$

The corresponding k-th step targeted MLE of ψ_0 is defined as $\psi_n^k = \Psi(Q_n^{*k}) \equiv \Psi(Q_{1n}^*, Q_{2n}^{*k})$. In this particular application, it follows that convergence occurs in one step so that $\psi_n = \Psi(Q_n^{*1})$.

The case-control weighted double robust estimating function for case control data is given by:

$$\begin{aligned} D_{q_0}(g^*, Q^*)(O) &= q_0 D^*(g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(g^*, Q^*)(M_1, W_2^j, A_2^j, 0), \end{aligned}$$

and the targeted MLE (g_n^*, Q_n^*) solves

$$0 = \sum_{i=1}^{n} D_{q_0}(g_n^*, Q_n^*)(O_i).$$

Statistical inference for ψ_n can be derived from the corresponding estimating equation $0 = \sum_{i=1}^{n} D(\psi_n, g_n^*, Q_n^*)(O_i)$ solved by the targeted MLE $\psi_n = \Psi(Q_n^*)$.

DOI: 10.2202/1557-4679.1114

2.7 Double robust locally efficient targeted MLE of treatment specific mean, causal relative risk and odds ratio for case control design I.

Let $\tilde{Q}_n^*$ be defined as a standard logistic regression fit ignoring the case control sampling. Subsequently, we map this into our estimator Q_{n,q_0}^* of Q_0^* by adding the intercept $\log c(q_0)$ to the log odds of $\tilde{Q}_n^*$.

We now construct an ϵ-fluctuation $Q_{n,q_0}^*(\epsilon)$ through the corresponding logistic regression fit $Q_{n,q_0}^*(Y \mid A, W)$ satisfying

$$\frac{d}{d\epsilon} \log Q_{n,q_0}^*(\epsilon) = D^*(Q_{n,q_0}^*, g_n^*),$$

where $D^*(Q^*, g^*)$ is the efficient influence curve of the bivariate parameter $(\Psi(Q^*)(0), \Psi(Q^*)(1))$ (i.e. EY_0, EY_1). This can be done by adding a two dimensional extension $\epsilon(I(A=1)/g_n^*(1 \mid W), I(A=0)/g_n^*(0 \mid W))$ to the log odds of the logistic regression fit Q_{n,q_0}^*.

Let

$$\epsilon_n = \arg\max_{\epsilon} \sum_i q_0 \log Q^*(W_{1i}, A_{1i}) + (1-q_0)\frac{1}{J}\sum_j \log(1 - Q^*(W_{2i}^j, A_{2i}^j))$$

be the case control weighted maximum likelihood estimator of ϵ, which can be fitted with standard logistic regression software again. The one-step targeted MLE of Q_0^* is now defined as $Q_n^* \equiv Q_{n,q_0}^*(\epsilon_n)$.

Since the update of the MLE Q_{n,q_0}^* only depends on g_n^* which does not change, it follows that this one-step targeted MLE Q_n^* already solves the case-control weighted efficient influence curve estimating equation:

$$\begin{aligned} 0 &= \sum_i q_0 D^*(Q_n^*, g_n^*)(W_{1i}, A_{1i}, 1) + (1-q_0)\frac{1}{J}\sum_j D^*(Q_n^*, g_n^*)(W_{2i}^j, A_{2i}^j, 0) \\ &\equiv \sum_i D_{q_0}(Q_n^*, g_n^*)(O_i), \end{aligned}$$

so that the generally prescribed iteration for targeted MLE is not needed.

The resulting targeted maximum likelihood estimator $\Psi(Q_n^*) = E_{Q_{W,n}^*} Q_n^*(a, W)$, with $Q_{W,n}^* = q_0 Q_{W_1,n}^* + (1-q_0) Q_{W_2,n}^*$ being the case control weighted empirical distribution of the covariate vector W, solves now the double robust estimating equation $0 = \sum_i D_{q_0}(Q_n^*, g_n^*, \Psi(Q_n^*))(O_i)$ (where we now use the estimating function representation of $D_{q_0}^*$), and is therefore a double robust estimator in the sense that it is consistent and asymptotically linear if either Q_n^* is consistent or g_n^* is consistent.

The same statistical properties are now established for the corresponding causal relative risks and odds ratios, where one uses that $Q_n^* = Q_{n,q_0}^*(\epsilon_n)$, just like Q_{n,q_0}^*, equals q_0 times a bounded estimator $Q_n^\#$ so that the standard error of this double robust targeted MLE is proportional to q_0 (divided by $\sqrt{n}$).

3 Case-control weighting of efficient procedure yields an efficient procedure for both case-control designs I and II.

In this section we state and show the remarkable nice result that assigning the case-control weights to the case-control sample and then applying an efficient procedure developed for prospective sampling actually yields an efficient procedure. These results are presented and derived for both case-control designs.

3.1 Case-control weighted mapping maps gradients into gradients.

Consider a target parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$. The class of all regular asymptotically linear estimators of $\Psi^*(P^*)$ at P^* can be characterized by their influence curves, and their influence curves constitute the set of gradients of the pathwise derivative of Ψ^* at P^* given a rich class of parametric fluctuations through P^*. In particular, an estimator is asymptotically efficient at P^* if and only if its influence curve equals the canonical gradient, that is, the unique gradient which is also an element of the tangent space generated by the scores of the class of parametric fluctuations. As a consequence of these general and powerful results an estimation problem is essentially characterized by the class of gradients and the canonical gradient. In particular, the class of gradients yields the class of wished estimating functions to construct double robust locally efficient estimators (van der Laan and Robins (2002)) and the canonical gradient provides the fundamental ingredient of the double robust locally efficient targeted maximum likelihood estimator.

This motivates us to identify the class of gradients, and, in particular, the canonical gradient, of the parameter Ψ^* in the case-control sampling model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ implied by the model $\mathcal{M}^*$ for the probability distribution P^* of interest and possible specification of dependence as identified by the η parameter, assuming that this parameter Ψ^* can be identified from case-control sampling.

DOI: 10.2202/1557-4679.1114

The following theorem establishes that the case-control weighting does provide a mapping from the set of all gradients of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$ into a set of gradients of $\Psi : \mathcal{M} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ at $P(P^*, \eta)$ in model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ for parameters Ψ^* which are identifiable from $P(P^*, \eta)$ (e.g. by being a function of q_0 or $\bar{q}_0(M)$). Since the class of all gradients of a parameter defined on a model represents the class of all possible influence curves of regular asymptotically linear estimators (see e.g, Bickel et al. (1993)), this result teaches us that the case-control weighting does map any estimation procedure developed for ψ_0^* based on prospective data into a corresponding estimation procedure based on case-control data, at least, from an asymptotic point of view.

In addition, since the case-control weighted mapping is 1-1, it also teaches us that it maps into a very rich set of estimation procedures for case-control data, if not all estimation procedures of interest: Indeed, we will show in the next subsections that the case-control weighted gradient mapping maps, in particular, into the optimal canonical gradient/efficient influence curve.

If the parameter of interest $\Psi^*(P^*)$ is only identified from $P = P(P^*, \eta)$ if q_0 and (for matched case-control designs) $\bar{q}_0$ is known, then one needs to define the parameter as a parameter indexed by the known q_0 and $\bar{q}_0(M)$: $\Psi^* = \Psi^*_{q_0}$.

We start with providing a useful definition of a gradient of a pathwise derivative.

Definition 2 *We define a gradient of pathwise derivative of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$ as a function $D^*(P^*)$ satisfying for each of the submodels $\{P^*_{S^*}(\epsilon) : \epsilon\} \subset \mathcal{M}^*$ through P^* at $\epsilon = 0$ with score S^* at $\epsilon = 0$ (within the class of submodels through P^* specified)*

$$\frac{d}{d\epsilon}\Psi^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = -\frac{d}{d\epsilon}P^*D(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0}.$$

Consider a parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ which is identified in model $\mathcal{M} = \{P = P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$, and corresponding parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$.

By the same definition of a gradient above, a gradient of the pathwise derivative of the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P = P(P^*, \eta)$ in model $\mathcal{M}$ is defined as a function $D(P^*, \eta)$ of O satisfying for each sub-model $\{P(P^*_{S^*}(\epsilon), \eta_{S_1}(\epsilon)) : \epsilon\} \subset \mathcal{M}$ implied by a submodel $\{P^*_{S^*}(\epsilon) : \epsilon\}$ through P^* and a nuisance submodel $\{\eta_{S_1}(\epsilon) : \epsilon\}$ through η indexed by S_1,

$$\Psi^*(P^*_{S^*}(\epsilon))|_{\epsilon=0} = -\frac{d}{d\epsilon}PD(P^*_{S^*}(\epsilon), \eta_{S_1}(\epsilon))\Big|_{\epsilon=0}.$$

Given this definition of a gradient we obtain the following theorem.

Theorem 3 *Given a $P^* \in \mathcal{M}^*$, a class of sub-models $\{P^*_{S^*}(\epsilon) : \epsilon\} \subset \mathcal{M}^*$ through P^* at $\epsilon = 0$ indexed by S^*, with score S^*, we have for each of these submodels*

$$\frac{d}{d\epsilon} P D_{q_0}(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = \frac{d}{d\epsilon} P^* D^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0}, \tag{2}$$

where it is assumed that the left and right derivative exist.

By (2) it follows that any gradient $D^*(P^*)$ of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$ is mapped into a gradient $D_{q_0}(P^*)$ of $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P = P(P^*, \eta)$ (for each η) in the model $\mathcal{M}$.

This last statement is an immediate consequence of (2) and the fact that $D_{q_0}(P^*)$ does only depend on $P = P(P^*, \eta)$ through P^* (and thus not through η), so that the derivatives along nuisance models $\{\eta(\epsilon) : \epsilon\}$ are zero, as required.

We now note that under extremely weak regularity conditions, the above definition of a gradient $D^*(P^*)$ of the pathwise derivative exactly agrees with the definition of a gradient of the pathwise derivative of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in efficiency theory (e.g., Bickel et al. (1993)), and similarly for Ψ. Namely, the equivalence follows if the second equality below holds (the first follows since $D^*(P^*) \in L_0^2(P^*)$): for the function $P^* \to D^*(P^*) \in L_0^2(P^*)$ and each submodel $\{P^*(\epsilon) : \epsilon\}$ (for each $P^* \in \mathcal{M}^*$) we have

$$\begin{aligned} \frac{1}{\epsilon} P^* D^*(P^*(\epsilon)) &= -\frac{1}{\epsilon} \int D^*(P^*(\epsilon)) \frac{dP^*(\epsilon) - dP^*}{dP^*(\epsilon)} dP^*(\epsilon) \\ &= -P^* D^*(P^*) S(P^*) + o(1), \end{aligned}$$

where $S(P^*)$ is the score $\frac{d}{d\epsilon} \log dP^*(\epsilon)/dP^*\Big|_{\epsilon=0}$ of the submodel $\{P^*(\epsilon) : \epsilon\}$.

For the interested reader, the following analogue theorem states the result in terms of the gradient of the pathwise derivative as in efficiency theory. That is, it provides the regularity condition under which we have that if $D^*(P^*)$ is a gradient of Ψ^* at P^*, then $D_{q_0}(P^*)$ is a gradient of the path-wise derivative of Ψ at $P(P^*, \eta)$.

Theorem 4 *Assume $\Psi : \mathcal{M} \to \mathbb{R}^d$ satisfies $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ for all $P^* \in \mathcal{M}^*$ and η.*

Assume $P^* \to D^*(P^*)$ is a gradient of the pathwise derivative of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in the sense that it satisfies for each member of a class of submodels $\{P^*_{S^*}(\epsilon) : \epsilon\}$ through $P^* \in \mathcal{M}^*$ at $\epsilon = 0$ with score S^*

$$\frac{d}{d\epsilon} \Psi^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = -\frac{d}{d\epsilon} P^* D^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0},$$

and the right-hand side equals $P^*D^*(P^*)S^*$, where it is assumed the derivative on the left and right-hand side exist.

Assume $P^* \rightarrow D_{q_0}(P^*)$ satisfies for each submodel $\{P(\epsilon) = P(P^*(\epsilon), \eta(\epsilon)) : \epsilon\} \subset \mathcal{M}$ through $P(P^*, \eta)$ at $\epsilon = 0$ (implied by the class of submodels $\{P^*_{S^*}(\epsilon)\}$ and $\{\eta_{S_1}(\epsilon)\}$) with score $S(P)$ that

$$-\frac{d}{d\epsilon} PD_{q_0}(P^*(\epsilon))\Big|_{\epsilon=0} = PD_{q_0}(P^*)S(P).$$

The latter is a regularity condition since

$$\begin{aligned} \frac{1}{\epsilon} PD_{q_0}(P^*(\epsilon)) &= -\frac{1}{\epsilon}\int D_{q_0}(P^*(\epsilon))\frac{dP(\epsilon) - dP}{dP(\epsilon)} dP(\epsilon) \\ &= -PD_{q_0}(P^*)S(P) + o(1), \end{aligned}$$

where $S(P)$ is the score $\frac{d}{d\epsilon}\log dP(\epsilon)/dP\Big|_{\epsilon=0}$ of the submodel $\{P(\epsilon) : \epsilon\}$.

Then, $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ is pathwise differentiable in the sense that for each of the submodels $\{P(\epsilon) = P(P^*(\epsilon), \eta(\epsilon)) : \epsilon\} \subset \mathcal{M}$ through $P(P^*, \eta)$ at $\epsilon = 0$ with score $S(P)$ we have

$$\frac{d}{d\epsilon}\Psi(P(\epsilon))\Big|_{\epsilon=0} = PD_{q_0}(P)S(P),$$

and $D_{q_0}(P)$ is a gradient of the pathwise derivative.

Thus, for each gradient $D^*(P^*)$ of the pathwise derivative of $\Psi^* : \mathcal{M}^* \rightarrow \mathbb{R}^d$ satisfying the above mentioned regularity conditions, the corresponding $D_{q_0}(P^*)$ is a gradient of the pathwise derivative of $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$.

Proof. We have

$$\begin{aligned} \frac{\Psi(P(\epsilon)) - \Psi(P)}{\epsilon} &= \frac{\Psi^*(P^*(\epsilon)) - \Psi^*(P^*)}{\epsilon} \\ &= -\frac{d}{d\epsilon} P^*D^*(P^*(\epsilon))\Big|_{\epsilon=0} + o(1) \\ &= -\frac{d}{d\epsilon} PD_{q_0}(P^*(\epsilon))\Big|_{\epsilon=0} + o(1) \\ &= PD_{q_0}(P^*)S(P) + o(1). \end{aligned}$$

This proves that $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ is pathwise differentiable at $P = P(P^*, \eta) \in \mathcal{M}$ and that $D_{q_0}(P^*)$ is a gradient of this pathwise derivative. □

Thus, the above result shows that each gradient $D^*(P^*)$ for $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ is mapped into a gradient $D_{q_0}(P^*)$ for $\Psi : \mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$. We note that this gradient mapping is not affected by the particular choice (i.e., model of dependence structure of case and control observations) of model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ compatible with $\mathcal{M}^*$. Thus, for example, for case-control design I, our mapping from gradients into gradients for model $\mathcal{M}$ is the same for the independence model assuming the case and controls are all independent as it is for a particular dependence model.

A particular case is that $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ is defined on a nonparametric model $\mathcal{M}^*$. In this case, there exists only one gradient for model $\mathcal{M}^*$ so that one just needs to determine the canonical gradient $D^*(P^*)$ of Ψ^* at P^* and map it into its case-control weighted version $D_{q_0}(P^*)$, which, by our results in the next section, equals the canonical gradient of Ψ at $P(P^*, \eta)$.

Remark. Since q_0 is a non-identifiable parameter for both case-control designs (so that knowledge of q_0 does not restrict the distribution of the data structure O), this implies that 1) for each gradient $D^*(P^*)$ for model $\mathcal{M}^*$, the corresponding $D_{q_0}(P^*)$ is a gradient in the model $\mathcal{M}$ *also including* the knowledge that q_0 is known (even if that knowledge was not included in $\mathcal{M}^*$), or, equivalently, the class of all gradients $\{D_h^*(P^*) : h\}$ at P^* for model $\mathcal{M}^*$ is mapped into a class $\{D_{h,q_0} : h\}$ of gradients at $P = P(P^*)$ for model $\mathcal{M}$ also including q_0 is known.

For matched case-control design II, if we define our parameter as $\Psi^*_{q_0}$, indexed by q_0 and $\bar{q}_0(M)$ (treating them as known and fixed), then the case-control weighting maps the class of all gradients of this parameter for model $\mathcal{M}^*$ into the class of gradients of this parameter for model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$. If the observed data model is the same with and without the restriction that $(q_0, \bar{q}_0(M))$ is known in the model $\mathcal{M}^*$, then the canonical gradient in the model $\mathcal{M}$ will be the same as the canonical gradient of the model also including the knowledge of $(q_0, \bar{q}_0(M))$.

DOI: 10.2202/1557-4679.1114

3.2 Independence models for case-control designs I and II to derive efficiency results.

We consider the independence model $\mathcal{M}$ so that $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$, where for case-control design I, we have

$$dP(P^*)(W_1, A_1, (W_2^j, A_2^j : j)) = dP^*(W_1, A_1 \mid Y = 1) \prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid Y = 0), \tag{3}$$

and, for case-control design II, we have

$$\begin{aligned} dP(P^*)(M_1, W_1, A_1, (M_1, W_2^j, A_2^j : j)) &= dP^*(M_1, W_1, A_1 \mid Y = 1) \\ &\prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \\ &= dP_M^*(M_1) dP^*(W_1, A_1 \mid M = M_1, Y = 1) \\ &\prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \end{aligned} \tag{4}$$

Our results immediately generalize to models $\mathcal{M}$ for which the densities of the distributions $P(P^*, \eta)$ factorize as

$$dP(P^*, \eta) = dP_1(P^*) dP_2(\eta),$$

where $dP_1(P^*)$ is given by the independence likelihood (3) or (4), and P^* and η are variation independent. This follows from the fact that such models the tangent space contains the tangent space of the independence model, and our proof of the wished result is based on showing that the case-control weighted efficient influence curve is a member of the tangent space and thereby equals the efficient influence curve for the model $\mathcal{M}$.

Our results in this section show that the case-control weighting of the canonical gradient for the prospective sampling model $\mathcal{M}^*$ yields the canonical gradient for the parameter of interest Ψ based on case-control sampling model $\mathcal{M}$. Our results rely on the assumption that (the typically very large/semiparametric) $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$ for case-control design I, and that $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and m varying over the support of the matching variable M.

As a consequence of our results, our proposed case-control weighted targeted maximum likelihood estimator for variable importance and causal effect

parameters, involving selecting estimators of Q_0^* and g_0^*, under appropriate regularity conditions guaranteeing the wished convergence of the standardized estimator to a normal limit distribution, is efficient if both of these estimators are consistent, and remains consistent if one of these estimators is consistent.

We note that the working-model to obtain the initial model based maximum likelihood estimators in our double robust targeted maximum likelihood estimator is obtained by modeling the factors of

$$dP^*(W, A, Y) = dP^*(W)dP^*(A \mid W)dP^*(Y \mid A, W),$$

which does thus not correspond with separate models for $dP^*(W, A \mid Y = \delta)$ as we "required" for the actual model $\mathcal{M}^*$ in order to make sure that the case-control weighted canonical gradient is a canonical gradient. In order to understand the rational of this discrepancy we provide the following explanation.

It happens to be that the efficient influence curve for our parameter of interest Ψ for an underlying model $\mathcal{M}^*$ identified by separate models for $P(W, A \mid Y = \delta)$ has a double robust representation in terms of Q_0^* and g_0^*, while it does not have a double robust representation w.r.t. to say $P(W, A \mid Y)$ or factors thereof. To fully exploit this double robust representation of the efficient influence curve of our parameter of interest, one should base estimation of the unknowns parameters of the efficient influence curve on the latter representation, and that is why we proposed our particular double robust locally efficient targeted maximum likelihood estimators.

Alternatively, we could use a targeted maximum likelihood estimator based on initial estimators based on working models for $P(W, A \mid Y = \delta)$, $\delta \in \{0, 1\}$: in this manner we would obtain generalized locally efficient double robust estimators where the double robustness is stated in terms of the models for Q_0^* and g_0^* *implied by* the models for $P(W, A \mid Y = \delta)$.

3.3 Case-control weighting of canonical gradient yields canonical gradient: Case Control Design I.

Firstly, we present the theorem for case-control design I.

Theorem 5 *Consider case-control design I. Assume that the model $\mathcal{M}^*$ allows independent variation of $P^*(W, A \mid Y = 1)$ and $P^*(W, A \mid Y = 0)$.*

Let $D^*(P^*)$ be the canonical gradient of the pathwise derivative $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$, let $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$ be the independence model defined by (3), and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ satisfy $\Psi(P(P^*)) = \Psi^*(P^*)$ for all

$P^* \in \mathcal{M}^*$. Assume the regularity conditions for $P^* \to D^*(P^*)$ of Theorem 4 apply so that it follows that Ψ is pathwise differentiable at P^* and $D_{q_0}(P^*)$ is a gradient of this pathwise derivative.

We have that $D_{q_0}(P^*)$ is the canonical gradient of the pathwise derivative of $\Psi : \mathcal{M} \to \mathbb{R}^d$.

We already knew that, if we set $D^*(P^*)$ equal to the canonical gradient (or any other gradient) of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$, then its case-control weighted version $D_{q_0}(P^*)$ is a gradient of $\Psi : \mathcal{M} \to \mathbb{R}^d$. The surprising and important extra result is that this $D_{q_0}(P^*)$ actually equals the canonical gradient. That is, for case-control design I, the case-control weighted gradient mapping does not only map gradients into gradients, it also maps the optimal canonical gradient for model $\mathcal{M}^*$ into the optimal canonical gradient for the observed data model $\mathcal{M}$ for case-control data.

Remark regarding q_0 known in model $\mathcal{M}^*$. Since q_0 is a non-identifiable parameter based on case-control sampling (design I), assuming q_0 is known in model $\mathcal{M}^*$ puts no restriction on the observed data model $\mathcal{M}$. As a consequence, the efficient influence curve for the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ is the same for the model $\mathcal{M}^*$ in which this quantity is known as it is in the model in which this quantity is unknown.

3.4 Example of efficient method for case-control design II based on stratified efficient method for case-control design I.

Before we present our general analogue result for case-control design II, it is helpful to consider an example for case-control design II. Consider the data structure $O^* = (M, W, A, Y) \sim P_0^*$ and let $\mathcal{M}^*$ be a nonparametric model. Consider case-control design II, in which our observed data $O = ((M_1, W_1, A_1), ((W_2^j, A_2^j) : j = 1, \ldots, J))$. Suppose we wish to estimate $\psi_0^* = E_0^* Y_1 = E_0^* E_0^*(Y \mid A = 1, M, W)$ and that $q_0(\delta \mid m) = \delta P_0^*(Y = 1 \mid M = m) + (1 - \delta) P_0^*(Y = 0 \mid M = m)$ is known. Recall that the efficient influence curve for this parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}$ in model $\mathcal{M}^*$ at P^* is given by $D^*(Q^*, g^*) - \psi^* = I(A = 1)/g^*(1 \mid M, W)(Y - Q^*(M, W, A)) + Q^*(M, W, 1) -$ *.

Consider the following general approach for estimation of ψ_0^* based on data generated by a case-control design II:

- Apply the case-control weighted targeted MLE for case-control design I to the subsample $\{i : M_{1i} = m\}$ to estimate the conditional version ${}^*_0(m) = E^*(Y_1 \mid M = m)$ of the parameter ψ_0^*. Thus this corresponds with weighting the cases with $q_0(1 \mid m) = P_0^*(Y = 1 \mid M = m)$ and the controls with $q_0(0 \mid m) = P_0^*(Y = 0 \mid M = m)$ and applying the standard prospective targeted MLE based on an initial estimator of $Q_0^*(m, a, w) = P_0^*(Y = 1 \mid m, a, w)$ and $g_0^*(a \mid m, w) = P_0^*(A = a \mid M = m, W = w)$. By our results for case-control design I, we know that this estimator yields a double robust locally efficient estimator of $\psi_0(m)$.

  This case-control weighted targeted maximum likelihood estimator of ${}_0(m)$ based on the subsample $\{i : M_{1i} = m\}$ solves the m-specific case-control weighted efficient influence curve equation $0 = P_n D^*_{m,q_0}(Q_n^*, g_n^*) - \Psi^*(Q_n^*)(m)$ and can thus be represented as

$$ {}_n(m) = \frac{\sum_i I(M_{1i} = m) D_{m,q_0}(Q_n^*, g_n^*)(O_i)}{\sum_i I(M_{1i} = m)}, \tag{5} $$

  where

$$ \begin{aligned} &D_{m,q_0}(Q^*, g^*)(O) = \\ &q_0(1 \mid m) \left\{ \tfrac{I(A_1=1)}{g_0^*(1|m,W_1)} (1 - Q^*(m, W_1, 1)) + Q^*(m, W_1, 1) \right\} \\ &+ \tfrac{q_0(0|m)}{J} \left\{ \tfrac{I(A_2^j=1)}{g^*(1|m,W_2^j)} (0 - Q^*(m, W_2^j, A_2^j, 1)) + Q^*(m, W_2^j, A_2^j, 1) \right\}. \end{aligned} $$

  The rational behind the consistency of this estimator $\psi_n(m)$ follows directly from the identity

$$ E(Y_1 \mid M = m) = \frac{E_0 D_{m,q_0}(Q_0^*, g_0^*)(O) I(M_1 = m)}{P_0(M_1 = m)}. $$

- Now, note that

$$ P_0^*(M = m) = P_0(M_1 = m) \frac{q_0}{q_0(1 \mid m)}. $$

  Thus, one maps $\psi_n(m)$ into an estimator of ψ_0 by averaging it w.r.t. to $q_0/q_0(1 \mid M_{1i}) P_n(M_1 = m)$:

$$ \begin{aligned} {}_n &= \sum_m \left\{ \frac{1}{n} \sum_{i=1}^n I(M_{1i} = m) \frac{q_0}{q_0(1 \mid M_{1i})} \right\} \ {}_n(m) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_{1i} = m) D_{m,q_0}(Q_n^*, g_n^*)(O_i), \end{aligned} $$

where we used (5).

Again, the rational of this estimator of ψ_0 follows immediately from the following derivation:

$$\begin{aligned}
&E_0 \textstyle\sum_m \frac{q_0}{q_0(1|m)} I(M_1 = m) D_{m,q_0}(Q_0^*, g_0^*) \\
&= E_0 \tfrac{q_0}{q_0(1|M_1)} D_{M_1,q_0}(Q_0^*, g_0^*) \\
&= E_0 \tfrac{q_0}{q_0(1|M_1)} \left\{ q_0(1 \mid M_1) D^*(M_1, W_1, A_1, 1) + \textstyle\sum_j \frac{q_0(0|M_1)}{J} D^*(M_1, W_2^j, A_2^j, 0) \right\} \\
&= E_0 q_0 D^*(M_1, W_1, A_1, 1) + \tfrac{\bar{q}_0(M_1)}{J} \textstyle\sum_j D^*(M_1, W_2^j, A_2^j, 0) \\
&= E_0^* Y_1,
\end{aligned}$$

where we suppressed the dependence of $D^* = D^*(Q^*, g^*)$ on Q^*, g^*.

- We conclude that this estimator ψ_n of ψ_0^* corresponds with solving our proposed case-control weighted efficient influence curve equation $P_n D_{q_0,\bar{q}_0} - = 0$, where

$$D_{q_0,\bar{q}_0}(O) = q_0 D^*(M_1, W_1, A_1, 1) + \frac{\bar{q}_0(M_1)}{J} \sum_j D^*(M_1, W_2^j, A_2^j, 0).$$

We conclude that this general approach for estimation of ψ_0^* of applying the case-control weighted targeted MLE $\psi_n(m)$ of case-control design I to the sub-sample $\{i : M_{1i} = m\}$ to estimate the analogue $\psi_0^*(m)$ of the parameter of interest ψ_0^* (i.e., the same function but now applied to the conditional $P_0^*(\cdot \mid M = m)$), and subsequently averaging $\psi_n(m)$ w.r.t. $q_0/q_0(1 \mid m) P_n(M_1 = m)$, corresponds with using our for case-control design II proposed case-control weighting $D_{q_0,\bar{q}_0}$ of the efficient influence curve D^* for model $\mathcal{M}^*$. This suggests that $D_{q_0,\bar{q}_0}$ is indeed also, just as we showed for case-control design I, the efficient influence curve. Our results below confirm this.

3.5 Case-control weighting of canonical gradient yields canonical gradient: Matched Case Control Design.

For case-control design II, we establish the same result.

Theorem 6 *Consider case-control design II. In this theorem we use the notation:* $D_{q_0,\bar{q}_0}(P^*) = q_0 D^*(P^*)(M_1, W_1, A_1, 1) + \frac{\bar{q}_0(M_1)}{J} \sum_j D^*(P^*)(M_1, W_2^j, A_2^j, 0)$.

Assume that the model $\mathcal{M}^*$ allows independent variation of $P^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and possible outcomes m of M under P_0^*.

Let $D^*(P^*)$ be the canonical gradient of the pathwise derivative $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$, let $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$ be the independence model

defined by (4), and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ *satisfy* $\Psi(P(P^*)) = \Psi^*(P^*)$ *for all* $P^* \in \mathcal{M}^*$.

Assume the regularity conditions for $P^* \to D^*(P^*)$ *of Theorem 4 apply so that it follows that* Ψ *is pathwise differentiable and* $D_{q_0,\bar{q}_0}(P^*)$ *is a gradient of this pathwise derivative at* $P(P^*) \in \mathcal{M}$.

Then, $D_{q_0,\bar{q}_0}(P^*)$ *is the canonical gradient of the pathwise derivative of* $\Psi : \mathcal{M} \to \mathbb{R}^d$.

3.6 Selecting the efficient influence curve of unrestricted target parameter.

In order to define an identifiable parameter $\Psi(P(P^*)) = \Psi^*(P^*)$ of the case-control data generating distribution, one often needs to define Ψ^* as indexed by the known q_0 and possibly $\bar{q}_0$ parameters. We denote such a parameter with $\Psi^*_{q_0} : \mathcal{M}^* \to \mathbb{R}^d$ to stress its dependence on these known fixed quantities. Our results above for case-control designs I and II above prove that if $D^*(P^*)$ is the canonical gradient of $\Psi^*_{q_0}$ at P^*, then the case-control weighted $D_{q_0}(P^*)$ is the canonical gradient of $\Psi : \mathcal{M} \to \mathbb{R}^d$, where $\Psi(P(P^*)) = \Psi^*_{q_0}(P^*)$ for all $P^* \in \mathcal{M}$. The following theorem shows that one can typically replace $D^*(P^*)$ by the canonical gradient of the path-wise derivative of the unrestricted $\Psi^*(P^*) = \Psi_{q(P^*)}(P^*)$.

Theorem 7 *Consider the two pathwise differentiable parameters* $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$ *indexed by a fixed* $r_0 = r(P^*_0)$ *(e..g, representing* q_0 *and* $\bar{q}_0$*), and a corresponding parameter* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ *defined as* $\Psi^*(P^*) = \Psi^*_{r(P^*)}(P^*)$. *Thus,* $\Psi^*_{r_0}(P^*_0) = \Psi^*(P_0)$.

Assume that for all the sub-models $\{P^*(\epsilon) : \epsilon\}$ *for which* $\frac{d}{d\epsilon} r(P^*(\epsilon))\big|_{\epsilon=0} = 0$, *we have*

$$\frac{d}{d\epsilon}\Psi^*(P^*(\epsilon))\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\Psi^*_{r_0}(P^*(\epsilon))\Big|_{\epsilon=0}.$$

Assume that the fixed parameter r_0 *in* $\Psi^*_{r_0}$ *is locally non-identifiable at* P^* *in the model* $\mathcal{M}$ *in the sense that the tangent space at* $P(P^*) \in \mathcal{M}$ *generated by the submodels* $\{P^*(\epsilon) : \epsilon\}$ *at* P^* *for which* $\frac{d}{d\epsilon} r(P^*(\epsilon))\big|_{\epsilon=0} = 0$ *equals the tangent space at* $P(P^*) \in \mathcal{M}$ *generated by all submodels used in definition of pathwise derivative of* $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$.

If the conditions of Theorem 5 or Theorem 6 apply for this choice $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$, *then we also have, if* $D^*(P^*)$ *is the canonical gradient of* Ψ^* *at* P^*, *then the case-control weighted* $D_{q_0}(P^*)$ *is the canonical gradient of* $\Psi : \mathcal{M} \to \mathbb{R}^d$.

DOI: 10.2202/1557-4679.1114

Proof. This result is shown as follows. Let D^* be the canonical gradient of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ and let D_1^* be the canonical gradient of $\Psi_{q_0}^* : \mathcal{M}^* \to \mathbb{R}^d$. As a consequence of the first assumption, we have for all scores S of all these submodels $P^*(\epsilon)$ not changing q_0 (in first order),

$$\langle D^*, S \rangle_{P^*} = \langle D_1^*, S \rangle_{P^*}.$$

So, if we restrict our class of sub-models at P^* in the definition of the path-wise derivative to these sub-models in $\mathcal{M}^*$ not varying r_0 (which globally corresponds with restricting $\mathcal{M}^*$ to all P^* with $r(P^*) = r_0$, but path-wise differentiability at P^* only depends on local thickness of model at P^*), then we have that the canonical gradient for the corresponding class of submodels for the observed data model is given by the case-control weighted $D_{q_0}(P^*)$ and the latter also equals the case-control weighted $D_{1q_0}(P^*)$. So under this restriction on the class of submodels through P^* we have equality of the two case-control weighted canonical gradients corresponding with D^* and D_1^*. Now, by using that this restriction on the class of submodels does not change the tangent space for the observed data models, and therefore does not affect the canonical gradient representation at $P(P^*)$ of the parameter Ψ in the observed data model $\mathcal{M}$. Thus this $D_{q_0}(P^*)$, which equals $D_{1q_0}(P^*)$, also equals the canonical gradient for the class of all submodels used in the actual definition of the pathwise derivative. This completes the proof of the theorem. □

Since q_0 is non-identifiable for case-control design I it follows that case-control weighting of the canonical gradient of the unrestricted parameter Ψ^* also yields the wished canonical gradient of Ψ. The same would apply for the matched case-control design, if enforcing the restriction $(q_0, q_0(1 \mid m) = P_0^*(Y = 1 \mid M = m))$ in $\mathcal{M}^*$ does not reduce the observed data tangent space, but this remains to be verified.

3.7 Proof of Theorems 5 and 6.

We already know that for both designs $D_{q_0}(P^*)$ (defined as $D_{q_0,1-q_0}(P^*)$ for design I and defined as $D_{q_0,\bar{q}_0}$ for design II) is a gradient of the pathwise derivative of Ψ at $P(P^*)$. Therefore, it remains to show that $D_{q_0}(P^*)$ is an element of the tangent space $T(P(P^*)) \subset L_0^2(P(P^*))$ defined as the closure of the linear span of the scores of each of the submodels $\{P(\epsilon) : \epsilon\}$ within the Hilbert space $L_0^2(P(P^*))$.

In the Appendix we have a separate section establishing these results for both designs, stating that if we select $D^*(P^*)$ as the canonical gradient of Ψ^* at P^* and the model $\mathcal{M}^*$ allows independent variation of $P(W, A \mid Y = \delta)$ for

Design I and independent variation of $P(W, A \mid M = m, Y = \delta)$ for Design II, then $D_{q_0}(P^*)$ is an element of the tangent space at $P(P^*)$ in the observed case-control data model $\mathcal{M}$.

Here we provide a summary of the proof for case-control design I in order to provide the reader with an understanding of these results.

Since $D^*(P^*)$ is a canonical gradient it equals a score $\frac{d}{d\epsilon} dP^*(\epsilon)/dP^*|_{\epsilon=0}$ for a particular submodel $\{P^*(\epsilon) : \epsilon\}$ at $\epsilon = 0$, or it can be arbitrarily well approximated by such a sequence of scores. We first consider the case that $D^*(P^*)$ is itself a score.

The tangent space under the independence model for a nonparametric model $\mathcal{M}^*$ is an orthogonal sum of the Hilbert space $T_1(P) = \{S_1(W_1, A_1) : S_1\}$ of functions of (W_1, A_1) with mean zero, and the Hilbert space $T_2(P) = \{\sum_j S_2(W_2^j, A_2^j) : S_2\}$ with $S_2(W_2^j, A_2^j)$ having mean zero, $j = 1, \ldots, J$. For an actual model $\mathcal{M}^*$ these two Hilbert spaces are replaced by sub-spaces spanned by the scores of the allowed sub-models $\{P^*(\epsilon) : \epsilon\}$ through P^*. That is, $T_1(P)$ consists of (and is generated by) functions $\frac{d}{d\epsilon}\frac{dP^*(\epsilon)}{dP^*}(W_1, A_1 \mid Y = 1)|_{\epsilon=0}$, and $T_2(P)$ consists of (and is generated by) functions $\sum_j \frac{d}{d\epsilon}\frac{dP^*(\epsilon)}{dP^*}(W_2^j, A_2^j \mid Y = 0)|_{\epsilon=0}$, $j = 1, \ldots, J$. We assumed that the marginal distributions $P^*(W, A \mid Y = 1)$ and $P^*(W, A \mid Y = 0)$ are independently varied by these submodels, so that indeed the tangent space is an orthogonal sum of $T_1(P)$ and $T_2(P)$.

For notational convenience, we introduce the notation $\epsilon_0 = 0$. Let $D^*(P^*) = \frac{d}{d\epsilon_0}\frac{dP^*(\epsilon_0)}{dP^*}(W, A, Y)$ be a score. Since q_0 is non-identifiable, we can assume that $p^*(\epsilon)(Y = 1) = q_0$ for all ϵ. It follows that

$$\begin{aligned} q_0 D^*(P^*)(W_1, A_1, 1) &= q_0 \frac{1}{p^*(W_1, A_1, 1)} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1, 1) \\ &= q_0 \frac{1}{p^*(W_1, A_1 \mid Y = 1) q_0} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1 \mid Y = 1) q_0 \\ &= q_0 \frac{1}{p^*(W_1, A_1 \mid Y = 1)} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1 \mid Y = 1) \\ &\in T_1(P^*), \end{aligned}$$

since the latter term equals q_0 times a score of $P(\epsilon)(W_1, A_1)$ at $\epsilon = 0$ (which in particular has mean zero).

DOI: 10.2202/1557-4679.1114

Again, using that $P^*(\epsilon)(Y=0) = 1-q_0$ for all ϵ,

$$\begin{aligned}
&(1-q_0)D^*(P^*)(W_2^j, A_2^j, 0) = (1-q_0)\frac{1}{p^*(W_2^j, A_2^j, 0)}\frac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j, 0)\\
&= (1-q_0)\frac{1}{p^*(W_2^j, A_2^j | Y=0)(1-q_0)}\frac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j \mid Y=0)p^*(\epsilon)(Y=0)\\
&= (1-q_0)\frac{1}{p^*(W_2^j, A_2^j | Y=0)}\frac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j \mid Y=0)\\
&\equiv (1-q_0)S_2(W_2^j, A_2^j),
\end{aligned}$$

where the latter term equals is $1-q_0$ times a score of $P(\epsilon)(W_2^j, A_2^j)$ at $\epsilon = 0$ (which, in particular, has mean zero). It follows that

$$\frac{(1-q_0)}{J}\sum_j D^*(P^*)(W_2^j, A_2^j, 0) = \frac{1-q_0}{J}\sum_j S_2(W_2^j, A_2^j) \in T_2(P(P^*)).$$

This proves that for case-control design I, if $D^*(P^*)$ is a score, then

$$D_{q_0}(P^*)(O) = q_0 D^*(P^*)(W_1, A_1) + \frac{1-q_0}{J}\sum_j D^*(P^*)(W_2^j, A_2^j)$$

is a score itself, and thus an element of the tangent space $T(P)$.

Suppose now that $D^*(P^*) = \lim_{m\to\infty} D_m^*(P^*) \in T^*(P^*)$, where $D_m(P^*) \in L_0^2(P^*)$ is a score. Then, for each m, we have $D_{mq_0}(P^*) \in L_0^2(P(P^*))$ is a score. To show that $D_{q_0}(P^*) \in L_0^2(P^*)$ is a score requires thus that the case-control mapping $D^* \to D_{q_0}$, as a mapping from $L_0^2(P^*)$ into $L_0^2(P(P^*))$ is continuous. This is trivially established. This proves that indeed $D_{q_0}(P^*)$ is an element of the tangent space $T(P(P^*))$. This completes the proof for case-control design I.

The proof for case-control design II is more delicate and provided in detail in the Appendix.

4 Summary, discussion and extensions.

We provide a generic approach for locally efficient estimation such as targeted maximum likelihood estimation of any parameter based on matched and unmatched case-control designs, which relies on specification of one or two non-identifiable parameters/scalars q_0 and, for matched case-control designs, $q_0(1 \mid m) = P_0^*(Y=1 \mid M=m)$.

These non-identifiable parameters could be known or they could be set in a sensitivity analysis, for example, in the case that these parameters are known to be contained in a particular interval. In the Appendix below we illustrate

how to handle the case that q_0 is replaced by a user supplied estimator based on an independent data set, and a standard error of this estimate of the true prevalence probability is provided.

Our approach is remarkably simple since it only requires weighting the cases by q_0 and the controls by $1 - q_0$ or $\bar{q}_0(M_1)$ and then applying a method developed for prospective sampling. Moreover, our approach has the remarkable convenient feature that applying the case-control weighting to an optimal method for the prospective sample results in an optimal method for independent and matched case-control designs.

We also showed how the case-control weighting for matched case-control designs corresponds with applying the case-control weighting for the standard unmatched case-control design for each sub-sample defined by a category for the matching variable to obtain the analogue conditional parameter, conditional on the matching variable category, and subsequently averaging these results over the matching variable categories to get the wished marginal parameter. This helps us to understand that our somewhat strange looking weights for the control observations in a matched case-control study are actually just as sensible as the much easier to understand weights for standard case-control designs.

In our accompanying technical report we worked out the case-control weighted targeted maximum likelihood estimators in a number of important applications involving estimation of variable importance and causal effect parameters. In addition, in our accompanying technical report we showed for both types of case-control designs how standard maximum likelihood logistic regression fits can be adjusted by using these known quantities to estimate conditional probabilities $P_0^*(Y = 1 \mid A, W)$ with a standard error which is proportional to q_0 divided by the square root of the sample size, so that the acquired precision results in stable estimators of such challenging parameters as relative risk and odds-ratios at $q_0 \approx 0$.

We believe that in many applications the marginal population proportion of cases, q_0, could be known, at least within close approximation, but it does require an effort to understand the target population the cases are sampled from. The literature supporting the use of q_0 in case-control studies goes back more than 50 years (See Cornfield (1951), Cornfield (1956)). Even 25 years ago, Greenland (1981) noted that "improvements in disease surveillance have produced more reliable estimates of disease incidence in many populations." Another relevant publication discussing the use of q_0 in case-control analysis is Benichou and Wacholder (1994).

In matched case-control studies in which one uses a matching variable with a large number of categories, then the value of the population proportion of

cases within each matching category might not be known. In that case, if the number of matching categories is large, a sensitivity analysis would likely be too cumbersome. On the other hand, even for such matched case-control samples, using the case-control weighting for design I might already provide an important bias reduction so that our methods only relying on q_0 will likely still provide a useful set of tools. Off course, this would require some validation that ignoring the matching does not cause severe bias.

During the design of a case-control study, we recommend to keep in mind that knowing these population proportion of cases for each matching category make the convenient and double robust efficient estimation of any causal effect and variable importance parameter possible (through the methods presented here) without restrictive assumptions such as the no-interaction assumption and parametric model form for conditional logistic regression models. This insight might help and motivate people to design case-control studies in which the required case-control weights are known or approximately known so that a sensitivity analysis is possible.

In addition, we note that the binary Y conditioned upon in the case-control sampling does not need to be an outcome of interest. For example, the random variable of interest might be a right-censored data structure $O^* = (W, A, \tilde{T} = \min(T, C))$, with T survival, C censoring, W covariates and A treatment, and in the case-control sampling we might condition upon a person having been observed to fail or not by time τ: $Y = I(\tilde{T} \leq \tau)$. In such an application the parameter of interest might be the causal effect of A on T.

To summarize, , by knowing q_0, one has available more efficient and more robust (i.e., double robust) targeted maximum likelihood estimators, targeting an identifiable parameter, and one does not have to restrict oneselves to odds-ratio parameters.

We now consider a few direct extensions and applications of our methodology.

Frequency matching. Frequency matching in case-control studies is typically defined as running a case-control design I within each strata $M = m$. In this case one can estimate any causal parameter $\psi_0(m)$ of the conditional distribution of O^*, given $M = m$, by assigning weights $q_0(1 \mid m)$ to the cases and $q_0(0 \mid m)/J$ to the corresponding J controls. Thus our methods for case-control design I can be applied to each strata $M = m$. In particular, this yields a locally efficient double robust targeted maximum likelihood estimator of $\psi_0(m)$ for each m. In order to estimate the marginal parameter ψ_0 one would need an estimate of the marginal distribution of M, which cannot be

identified based on knowing $q_0(1 \mid m)$ only, so that other knowledge will be needed such as the marginal population distribution of M. Either way, one can always estimate causal parameters such as $E(Y_a \mid M = m)$ for each m or the corresponding variable importance measure. If the number of categories of the matching variable is large, then a sensible strategy for estimation of $\psi_0(m)$ is to assume a model $\psi_0(m) = f(m \mid \beta_0)$ and obtain a pooled locally efficient targeted maximum likelihood of β_0 based on all observations.

Pair matching. Pair matching in case-control studies is typically described as, for each matching category, sample a case and a set of controls. So this description agrees with frequency matching except that the number of categories can be very large. Therefore, we should now always assume a model $_0(m) = f(m \mid \beta_0)$ and obtain a pooled locally efficient targeted maximum likelihood of β_0 based on all observations.

Without the knowledge of $q_0(1 \mid m)$, one would use conditional logistic regression models, and, as noted in Jewell (2006) page 258, these methods do not allow estimation of the association of M with Y, while if one knows the population proportion $q_0(1 \mid m)$ we can estimate every parameter of the population distribution, conditional on $M = m$.

Counter matching. Finally, another type of matching in case-control studies is called counter-matching, which involves sampling a control with an exposure (maximally) different from the exposure of the case. Formally, we can define this sampling scheme as follows. The observation $O = ((M_1, Z_1), (M_2, Z_2))$ on each experimental unit is generated as 1) sample (M_1, Z_1) from the conditional distribution of (M, Z), given $Y = 1$, and 2) sample (M_2, Z_2) from the conditional distribution of (M, Z), given $M = m^*(M_1)$ and $Y = 0$, where $m^*(m)$ maps a particular outcome m into a counter-match $m^*(m)$ in the outcome space for M. Similarly, this is defined for the case that one samples J controls counter-matched to the case. The population distribution of interest is the distribution P_0^* of $O^* = (M, Z, Y)$ and we are concerned with estimation of a particular parameter ψ_0^* of this distribution P_0^* based on a counter-matched case-control sample $O_1, \ldots, O_n$. In this case, given that $D^*(M, Z, Y)$ satisfies $P_0^* D^* = 0$, we have

$$E_0 D_{q_0, \bar{q}_0^*}(O) = 0,$$

where the case-control weighted version of D^* is defined as

$$D_{q_0, \bar{q}_0}(O) = q_0 D^*(M_1, Z_1, 1) + \bar{q}_0^*(M) D^*(m^*(M_1), Z_2, 0),$$

DOI: 10.2202/1557-4679.1114

with

$$\bar{q}_0^*(m) = (1 - q_0)\frac{P_0^*(M = m^*(m) \mid Y = 0)}{P_0^*(M = m \mid Y = 1)}.$$

Note that if $m^*(m) = m$ is the identity function, then indeed $\bar{q}_0^* = \bar{q}_0$. The non-identifiable component of the control-weight $\bar{q}_0^*$ is $P_0^*(M = m^*(m), Y = 0)$, or, assuming q_0 is known, $P_0^*(M = m^*(m) \mid Y = 0)$, while the denominator $P_0^*(M = m \mid Y = 1) = P_0(M_1 = m)$ can be empirically estimated. Since in many applications the control observations are relatively easily accessible, one might use a separate sample of controls to estimate these proportions $P_0^*(M = \cdot \mid Y = 0)$ having a certain value for the (counter-)matching variable M among the controls. So under the condition that these weights $q_0, \bar{q}_0^*$ are known (or set in a sensitivity analysis), our results in this article can be applied to counter-matched case-control designs by just replacing $\bar{q}_0$ by $\bar{q}_0^*$.

Propensity score matching design. A commonly used design is the following. One samples from the units that received treatment. For each treated unit, one finds a matched non-treated unit, where the matching is done based on a fit of the so called propensity score. The goal of this design is to create a sample in which the confounders are reasonably balanced between the treated and untreated units. This design can formally be described as follows. The random variable of interest is $O^* = (W, A, Y) \sim P_0^*$, and one is typically concerned with estimation of a causal effect such as $E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\}$. Let $M \equiv \Pi^*(W)$ be a summary measure of W which is supposedly an approximation of the propensity score $\Pi_0^*(W) = P_0(A = 1 \mid W)$ (e.g., estimated from external data). One samples $(M_1 = \Pi^*(W_1), W_1, Y_1)$ from the conditional distribution of (W, Y), given $A = 1$, and one samples one or more $(M_2 = \Pi^*(W_2), W_2, Y_2)$ from the conditional distribution of (W, Y), given $M = M_1$and $A = 0$.

One now wishes to use n i.i.d. observations on the observed experimental unit $O = ((W_1, Y_1), (W_{2j}, Y_{2j} : j))$ representing a treated unit and one or more propensity score matched untreated units to estimate the causal parameter of interest.

Notice that we can immediately apply the methodology presented in this article by defining the Y as the A and the matching variable M is playing the role of $\Pi^*(W)$. As a consequence, one can use any method developed for sampling from (W, A, Y) by using our "case control" weights $q_0 = P_0^*(A = 1)$ for the treated units, and $\bar{q}_0(W) = q_0 \frac{P_0^*(A=0|M)}{P_0^*(A=1|M)}$ for the untreated units. Thus, to correct for the biased sampling one will need to know the actual true treatment mechanism/propensity score $P_0^*(A = 1 \mid W)$. Thus, under

the assumption that this propensity score is known or can be estimated based on an external data source, one can apply any method for estimation of the wished causal effect for standard sampling by applying these weights to the treated and untreated units. Off course, for the sake of statistical inference and model selection (say, based on cross-validation) one should respect the fact that the independent and identically distributed observations are $O_1, \ldots, O_n$, and not the treated and untreated units.

General biased sampling. Finally, we like to discuss the implications of the proposed optimal case-control weighting for general biased sampling models with known probabilities for the conditioning events, where optimal refers to the fact that the case-control weighting maps an efficient procedure for an unbiased sample into an efficient procedure for the biased sample. The following generalization of our method for case-control design I applies to general biased sampling. Consider a particular target probability distribution P_0^* representing the unbiased sampling distribution and its corresponding random variable $O^* \sim P_0^*$. Suppose now that the outcome space for the random variable O^* is partitioned by a union of events $\mathcal{A}_j$, $j = 1, \ldots, J$: i.e. $Pr(O^* \in \cup_j \mathcal{A}_j) = 1$ and the sets $\mathcal{A}_j$ are pairwise disjoint. Let the experimental unit for the observed data be $(O_1, \ldots, O_J)$, where $O_j \sim O^* \mid O^* \in \mathcal{A}_j$ is a draw from the conditional distribution, given $O^* \in \mathcal{A}_j$, $j = 1, \ldots, J$. For simplicity, we enforced here equal number of draws, but this can be generalized to having different number of draws from each conditional distribution. Let $q_0(j) = P_0^*(O^* \in \mathcal{A}_j) \in (0, 1)$ and suppose these probabilities are known. Weighting observation O_j with $q_0(j)$ for $j = 1, \ldots, J$, and applying a method developed for the unbiased sample will yield valid estimators. We also conjecture that under appropriate similar conditions as we assumed for case-control sampling, this weighting will be optimal in the sense that assigning these weights to an efficient estimation procedure for i.i.d. samples of P_0^* will yield an efficient estimation procedure based on the biased sampling model. Given our interpretation of case-control weighting for matched case-control sampling in terms of case-control weighting for standard case-control studies conditional on the matching category, we suggest that weighting for matched case-control sampling can be generalized to matched biased sampling in general (say matched on a draw M_1 from the first biased sampling distribution).

Another commonly employed study is a case-control sample nested within a cohort. In addition, it is then common that one collects additional information on the case-control sample relative to the information collected in the original cohort sample. Our results are not covering this important problem for which

a rich literature exist (see e.g., Robins et al. (1994)).

Appendix: Incorporating variability/uncertainty in the user supplied prevalence probability q_0.

In this section we wish to illustrate that our general case-control weighted estimation methodology directly generalizes to the case that q_0 is replaced by an estimate $\hat{q}$ (based on an independent sample) with a user supplied standard error σ. For the sake of illustration, consider the independent case-control design and let $D_{q_0}(O \mid \psi) = q_0 D(W_1, A_1, 1 \mid \psi) + (1 - q_0)/J \sum_j D(W_0^j, A_0^j, 0 \mid)$ be a case-control weighted estimating function applied to an estimating function $D(O^* \mid \psi)$ for the parameter of interest $\psi_0 = \Psi(P_0^*)$ of the target distribution P_0^*. Let the case-control weighted estimator $\hat{\Psi}(q_0, P_n)$ be defined as a solution of the estimating equation

$$0 = P_n D_{q_0}(O \mid \psi) = \frac{1}{n} \sum_{i=1}^{n} D_{q_0}(O_i \mid \psi),$$

where P_n denotes the empirical distribution.

The case-control weighted estimator ψ_n based on $\hat{q}$ of ψ_0 can now be represented as $\hat{\Psi}(\hat{q}, P_n)$. Under regularity conditions, the estimator $\hat{\Psi}(q_0, P_n)$ (as consider in our article) using the true prevalence probability q_0 is asymptotically linear with influence curve $IC_0 = -\frac{d}{d\psi_0} P_0 D_{q_0}(\psi_0)^{-1} D_{q_0}(\psi_0)$, using short-hand notation. The actual estimator $\hat{\Psi}(\hat{q}, P_n)$ can now be decomposed as

$$\begin{aligned} \hat{\Psi}(\hat{q}, P_n) - \psi_0 &= \hat{\Psi}(\hat{q}, P_n) - \hat{\Psi}(\hat{q}, P_0) + \hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0) \\ &\approx \hat{\Psi}(q_0, P_n) - \hat{\Psi}(q_0, P_0) + \hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0), \end{aligned}$$

where the approximation involves a second order term of $\hat{q} - q_0$ and $P_n - P_0$. The first difference equals $(P_n - P_0)IC_0 + o_P(1/\sqrt{n})$ and is thus asymptotically normally distribution with mean zero and covariance matrix $\Sigma_0 = E_0 IC_0 IC_0^\top$. The second difference is independent of this first asymptotically normal term and, by the delta-method, can be approximated by $\hat{q} - q_0$ times the gradient a_0 of $q \to \hat{\Psi}(q, P_0)$:

$$\hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0) = (\hat{q} - q_0) \frac{d}{dq_0} \hat{\Psi}(q_0, P_0) = (\hat{q} - q_0) a_0.$$

Thus, this term behaves as a normally distributed vector with mean zero and variance elements $\sigma^2 a_0$, where $a_0 = \frac{d}{dq_0}\hat{\Psi}(q_0, P_0)$. We can conclude that our standardized estimator $\sqrt{n}(\hat{\Psi}(\hat{q}, P_n) - \psi_0)$ converges in distribution to

$$N(0, \Sigma + \sigma^2 a_0 a_0^\top),$$

where $\Sigma = E_0 IC_0(O) IC_0^\top(O)$ is the covariance matrix of the normal limit distribution of the estimator $\hat{\Psi}(q_0, P_n)$ based on the known prevalence probability.

In general, this general template shows that we can incorporate the standard error σ of a user supplied estimate $\hat{q}$ by simply adding the matrix $\sigma^2 a_0 a_0^\top$ to the covariance matrix of our case-control weighted estimator $\hat{\Psi}(\hat{q}, P_n)$ we would use if $\hat{q}$ is treated as known, where a_0 is the gradient of $q \to \hat{\Psi}(q, P_0)$ at q_0.

For the sake of concreteness, we will now provide an expression of the gradient a_0 of the derivative of $q \to \hat{\Psi}(q, P_0)$ at $q = q_0$ in the above setting. Note that $\hat{\Psi}(q, P_0)$ is defined as the solution in ψ of $H_0(q, \psi) = P_0 D_{q_0}(\psi) = 0$. By the implicit function theorem, this shows that the gradient of $q \to \hat{\Psi}(q, P_0)$ is given by:

$$\begin{aligned} a_0 &= -\frac{d}{d\psi_0} H_0(q_0, \psi_0)^{-1} \frac{d}{dq_0} H_0(q_0, \psi_0) \\ &= -\frac{d}{d\psi_0} H_0(q_0, \psi_0)^{-1} P_0(D_1 - D_0), \end{aligned}$$

where we defined $D_1(O) = D(1, W_1, A_1)$ and $D_0(O) = \frac{1}{J}\sum_j D(0, W_0^j, A_0^j)$. One can estimate a_0 by replacing the expectations by empirical means, and thereby construct confidence intervals and p-values based on $\Sigma_n + \sigma^2 a_n a_n^\top$, where Σ_n is an estimator of the covariance matrix Σ_0 and a_n is the estimator of a_0.

Appendix: Extension to case-control incidence density sampling.

An alternative commonly employed case-control sampling design involves regular case-control sampling from a population at risk at time t, where the outcome is now defined at time t, across various time points t (see e.g., Rothman and Greenland (1998)). Such designs can be carried out at only a few discrete time points or they could evolve in continuous time.

For example, one might sample breast cancer cases and controls in year 2000 among the population at risk of breast cancer, and one would repeat

DOI: 10.2202/1557-4679.1114

such a case-control sample at years 2001 and 2002. Note that the outcome is now different depending on the year one samples, since being a case in the case-control sample at year (e.g.) 2000 requires being diagnosed with breast cancer in year 2000. Another type of example would be to sample one or more controls at the time a case occurs among the subjects at risk right before the case occurred.

One issue with this kind of case-control sampling is that the sampling population might change over time due to an influx of new subjects over time, so that the change in sampling population over time cannot only be modeled by censoring and the occurrence of failures within a well defined target population at the first time point. Alternatively, one defines a target population at the first time point and one samples cases and controls at time t among the subjects in this target population that are still at risk right before time t (i.e., the subjects that have not failed or been censored, yet), thereby ignoring any possibly influx of subjects over time.

We now wish to discuss some possible applications of our case-control weighting methodology to these types of case-control sampling designs. Firstly, the most straightforward and direct application is to treat the case-control sample at time t as a separate case-control sample and immediately apply our case-control weighting to estimate any parameter of the population distribution one samples from at time t. Of course, this requires a large enough case-control sample at each time point t so that these t-specific parameters are estimated at a reasonable precision. Note also that the knowledge of the case-control weights now requires knowing the marginal probability of being a case for the sampling population at time t, at each of the sampling times t. If one is willing to assume that these t-specific parameters (e.g. causal effect of a treatment on outcome) follow a parametric trend in t, then one can pool all the t-specific estimates to obtain a smoother estimation procedure that might result in significant gains in variance. For example, maybe it is appropriate to believe that the population is stationary in time t, that is, somehow the influx of new subjects and loss of existing subjects due to censoring or failure balances out so that the sampling population at time t does not change over time. In that case, one might assume that the t-specific parameters are constant in t.

We now wish to consider how we might generally apply pooling across time while using our case-control weighting to handle such incidence density sampling designs. Here, we will focus on a single target population so that one is concerned with estimation of a single well defined parameter of a target population of interest.

Consider the case that the outcome of interest is a time till event T. For

notational convenience, we will assume that T is discrete on time points $t = 0, 1, \ldots, \tau$. Suppose that in a prospective sample one would observe $O^* = (\tilde{T} = \min(C, T), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T}))$, where C is a right-censoring time and $\bar{X}(s)$ denotes the history up till time s of the time dependent process $t \to X(t)$: $\bar{X}(s) = (X(u) : u \leq s)$, where $X(t)$ includes the indicator $dY(t) \equiv I(T = t)$ of the failure time event at time t. Let P_0^* denote the probability distribution of this right-censored data structure O^*. Suppose that the parameter of interest is $\Psi(P_0^*)$ which will typically represent a parameter of the full data distribution of X such as a causal effect of a treatment A assigned at time 0 on the time till event T.

In the case that the outcome is a time till a *rare* event one might employ a so called incidence density case-control sampling design. That is, at time t, among the population at risk defined by all the individuals with $R(t) = I(\tilde{T} \geq t) = 1$, one samples a case from the conditional distribution of O^*, given $dY(t) = 1$ and $R(t) = 1$, and one samples one or more controls from the conditional distribution of O^*, given $dY(t) = 0$ and $R(t) = 1$. Note that one can replace $dY(t)$ by the observed data quantity $dY(t) = I(\tilde{T} = t, \Delta = 1)$. Let's denote the observed data structure sampled at time t, consisting of a case and one or more controls, as

$$O_t = (O_{1t}, O_{0tj},\ j = 1, \ldots, J),$$

where O_{1t} denotes the data structure on the case and O_{0tj} denotes the data structure on the j-th control. Suppose one samples $n(t)$ i.i.d observations of O_t at time t, $t = 0, \ldots, \tau$, resulting in a total sample O_{ti}, $i = 1, \ldots, n(t)$, $t = 0, \ldots, \tau$.

Let $R(t)D(t, O^*)$ be an estimating function or loss function for the prospectively sampled unit O^*, $t = 0, \ldots, \tau$. An estimating function or loss function based on sampling O^* itself can always be represented as $\sum_t R(t)D(t, \bar{O}^*(t))$, where $\bar{O}^*(t) = \bar{X}(\min(t, C))$ denotes the observed history up till time t, which is assumed to include the censoring event if it occurs before time t. Specifically, we have

$$\begin{aligned} D(O^*) &= \sum_t E(D \mid \bar{X}(\min(t, C))) - E(D \mid \bar{X}(\min(t-1, C))) \\ &= \sum_t R(t) \left\{ E(D \mid \bar{X}(\min(t, C)) - E(D \mid \bar{X}(\min(t-1, C))) \right\}, \end{aligned}$$

where $\bar{X}(\min(t, C))$ represents the history one observes up till time t, and thus it is assumed that $\bar{X}(\min(t, C))$ also includes observing the censoring event time C if C occurs before time t

The following lemma shows how the case-control weighting can be applied to this t-specific estimating function of O^* which typically represents just one

term $R(t)D(t,O^*)$ of the full estimating function $D(O^*) = \sum_t R(t)D(t,O^*)$ one would use if one would sample O^* prospectively.

Lemma 1 *Define*

$$D_{q_0}(t,O_t) \equiv q_0(t)D(t,O_{1t}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,O_{0tj}),$$

where

$$\begin{aligned} q_0(t) &\equiv P_0^*(dY(t)=1, R(t)=1) \\ \bar{q}_0(t) &\equiv P_0^*(dY(t)=0, R(t)=1). \end{aligned}$$

We have

$$E_0 D_{q_0}(t,O_t) = E_0^* R(t)D(t,O^*).$$

In particular, if we redefine $q_0(t) = P(dY(t)=1 \mid R(t)=1)$ *and* $\bar{q}_0(t) = 1-q_0(t)$*, then*

$$E_0 D_{q_0}(t,O_t)P_0^*(R(t)=1) = E_0^* R(t)D(t,O^*).$$

If censoring is non-informative, then the weights $q_0(t) = P_0^*(dY(t)=1 \mid R(t)=1) = P_0^*(dY(t)=1 \mid T \geq t)$ reduce to the marginal hazard of T at time t. Thus, if censoring is non-informative, then this case-control weighting would require knowing the marginal failure time distribution of T.

Proof of Lemma. We have

$$\begin{aligned} E_0 D_t(O_t) &= E_0 q_0(t)D(t,1,O_1^*) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,0,O_{0j}^*) \\ &= \int D(t,1,O_1^*)q_0(t)P_0^*(O^* \mid dY(t)=1, R(t)=1) \\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\int D(t,0,O^*)\bar{q}_0(t)P_0^*(O^* \mid dY(t)=0, R(t)=1) \\ &= \int_{O^*} D(t,1,O^*)P_0^*(O^*, dY(t)=1, R(t)=1) \\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\int_{O^*} D(t,0,O^*)P_0^*(O^*, dY(t)=0, R(t)=1) \\ &= \int_{O^*,R(t)} R(t)D(t,1,O^*)P_0^*(O^*, dY(t)=1, R(t)) \\ &\quad + \int_{O^*,R(t)} R(t)D(t,0,O^*)P_0^*(O^*, dY(t)=0, R(t)) \\ &= \int_{O^*,dY(t),R(t)} R(t)D(t,dY(t),O^*)P_0^*(O^*, dY(t), R(t)). \end{aligned}$$

This proves the lemma. □

Even though one only applies the t-specific component $R(t)D(t,O^*)$ of the full estimating function to the case, the following lemma shows that one can often use the control observation sampled at time t for the later time point estimating functions without any need for weighting or coupling them to the case sampled at time t.

Lemma 2 *Assume* $E_0(D(s,O^*) \mid R(s)=1)=0$ *for all* s. *Given a* t, *for* $s>t$, *we have for the control observations* O_{0t}

$$E_0R(s)D(s,O_{0t})=0$$

Proof. We have for $s>t$

$$\begin{aligned}
&E_0R(s)D(s,O_{0t}) = E_0(R(s)D(s,O^*) \mid R(t)=1,dY(t)=0)\\
&= E_0(E_0(R(s)D(s,O^*) \mid R(s),R(t)=1,dY(t)=0) \mid R(t)=1,dY(t)=0)\\
&= E_0(P_0(R(s)=1 \mid R(t)=1,dY(t)=0)\\
&\times E_0(R(s)D(s,O^*) \mid R(s)=1,R(t)=1,dY(t)=0) \mid R(t)=1,dY(t)=0)\\
&= P_0(R(s)=1 \mid R(t)=1,dY(t)=0)\\
&\times E_0(E_0(R(s)D(s,O^*) \mid R(s)=1) \mid R(t)=1,dY(t)=0)\\
&= P_0(R(s)=1 \mid R(t)=1,dY(t)=0)E_0(D(s,O^*) \mid R(s)=1)\\
&= 0.\square
\end{aligned}$$

Thus, given an estimating function $D(O^* \mid \psi)=\sum_t R(t)D(t,O^* \mid \psi)$ for the parameter ψ_0^* based on sampling from P_0^*, an estimating equation for the total sample from the actual biased sampling data generating distribution P_0 can now be constructed as:

$$\begin{aligned}
0 \;=\; &\sum_t \sum_{i=1}^{n(t)} q_0(t)D(t,O_{1ti}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,O_{0tji} \mid \psi)\\
&+\sum_t \sum_{i=1}^{n(t)} \sum_{s>t} \frac{1}{J}\sum_j R(s)D(s,O_{0sji} \mid \psi).
\end{aligned}$$

The last term represents the non-coupled contributions of the control observations at time points after the time point t at which the control unit was sampled. Here $q_0(t)=P(dY(t)=1 \mid R(t)=1)$ and $\bar{q}_0(t)=P(dY(t)=0 \mid R(t)=0)=1-q_0(t)$. If the estimating function is indexed by nuisance parameters, then these need to be estimated.

DOI: 10.2202/1557-4679.1114

Not conditioning on being at risk. In the above form of incidence density sampling, sampling a case corresponds with conditioning on a subject being at risk and being a true case at time t (i.e., a failure at time t). In the following lemma we employ the same design but in which sampling a case corresponds with only conditioning on having an observed event at time t, and thus not conditioning on being at risk at time t. The advantage of this type of design is that one can now case-control weight the complete estimating function $D(O^*) = \sum_t R(t)D(t, O^*)$ one would use in prospective/unbiased sampling of O^*. The lemma provides the correct case control weighting and it is now a direct corollary of our case-control weighting results established in this article.

Lemma 3 *Let $dY(t) = I(\tilde{T} = t, \Delta = 1)$. Let O_{1t} be a draw from conditional distribution of O^*, given $dY(t) = 1$, and let O_{0tj} be i.i.d draws from the conditional distribution of O^*, given $dY(t) = 0$, $j = 1, \ldots, J$. Let $O_t = (O_{1t}, (O_{0tj} : j))$ be the total observation consisting of a case and J controls.*

Given a function $O^* \rightarrow D(O^*)$, define

$$D_{q_0}(t, O_t) \equiv q_0(t)D(O_{1t}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(O_{0tj}),$$

where

$$\begin{aligned} q_0(t) &\equiv P_0(dY(t) = 1) \\ \bar{q}_0(t) &\equiv P_0(dY(t) = 0). \end{aligned}$$

We have

$$E_0 D_{q_0}(t, O_t) = E_0^* D(O^*).$$

Thus, given an estimating function $D(O^* \mid \psi) = \sum_t R(t)D(t, O^* \mid \psi)$ for the parameter ψ_0^* based on sampling from P_0^*, an estimating equation for the total sample from the actual biased sampling data generating distribution P_0 can now be constructed as

$$0 = \sum_t \sum_{i=1}^{n(t)} q_0(t)D(O_{1ti} \mid \psi) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(O_{0tji} \mid \psi),$$

$q_0(t) = P(dY(t) = 1)$ and $\bar{q}_0(t) = P(dY(t) = 0) = 1 - q_0(t)$. If the estimating function is indexed by nuisance parameters, then these need to be estimated.

If there is censoring, then, even if censoring is independent, $q_0(t)$ is also a function of the censoring mechanism for the prospective data structure O^*,

which might be viewed as a disadvantage of such weights. Again, we note that the case-control weighting requires knowing these marginal incidence probabilities $q_0(t)$ across all time points t to which one applies the case control sampling.

Matched case-control incidence density sampling. Since the weights of our lemmas are directly implied by our case-control weights used in this article, it is also clear how we can generalize the above lemmas to matched case-control incidence density sampling in which one matches the controls by also conditioning on a matching variable being equal to the matching variable of the case.

An example: Estimation of conditional hazards based on incidence density case-control sampling. Consider a target population of individuals, and suppose that the data structure O^* on a sampled individual consists of baseline covariates W, a treatment variable A, and a right-censored survival time T, so that $O^* = (W, A, \tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T)$. Suppose we are concerned with estimation of an intensity $E(dY(t) \mid \bar{F}(t), A, W)$ of the counting process $Y(t) = I(\tilde{T} \leq t, \Delta = 1)$ w.r.t to history $\bar{F}(t), A, W$, where $\bar{F}(t)$ represents the failure and censoring history up till time t. If censoring is conditionally independent of T given A, W, then this intensity equals $I(\tilde{T} \geq t)E(I(T \in dt) \mid T \geq t, A, W)$, that is, it equals the conditional hazard of T, given A, W. It is common to assume a Cox-proportional hazards or logistic regression model, depending on T being continuous (or finely discrete) or discrete. For the sake of illustration, let's consider a parametric model $\alpha_\beta(t \mid \bar{F}(t), A, W)$ for this intensity $E(dY(t) \mid \bar{F}(t), A, W)$ indexed by a finite dimensional parameter β. Under sampling from O^* it is known how to construct a good estimator of β_0. In particular one can apply maximum likelihood estimation where the likelihood for a single observation O^* is given by $\prod_t P_\beta(dY(t) \mid W, A, \bar{F}(t))$ in which

$$P_\beta(dY(t) \mid W, A, \bar{F}(t)) = \\ \alpha_\beta(t \mid W, A, \bar{F}(t))^{dY(t)}(1 - \alpha_\beta(dY(t) \mid W, A, \bar{F}(t))^{1-dY(t)}$$

represents the Bernoulli likelihood corresponding with the model α_β. Let $R(t)D_\beta(t, dY(t), W, A, \bar{F}(t))$ be defined as t-specific term $R(t) \log P_\beta(dY(t) \mid W, A, \bar{F}(t))$ of the log-likelihood $\sum_t R(t) \log P_\beta(dY(t) \mid W, A, \bar{F}(t))$ of O^*.

Consider now a case-control incidence density sampling design in which at time t one samples a case from the conditional distribution of O^*, given that $dY(t) = 1, R(t) = 1$, and one or more controls from the conditional

DOI: 10.2202/1557-4679.1114

distribution of O^*, given that $dY(t) = 0$, $R(t) = 1$. Let $O_t = (O_{1t}, (O_{0tj} : j))$ denote the coupled case and control observations. As above, let $n(t)$ denote the number of such observations one samples at time t: O_{ti}, $i = 1, \ldots, n(t)$. For notational convenience, let's assume that one samples a single control for each case: i.e., $J = 1$. As case-control weighted log-likelihood, augmented with the control observation contributions for later time points, we obtain:

$$\begin{aligned} L_n(\beta) &= \sum_t \sum_{i=1}^{n(t)} q_0(t) D_\beta(t, 1, W_{1ti}, A_{1ti}, \bar{F}_{1ti}(t)) \\ &\quad + \bar{q}_0(t) D_\beta(t, 0, W_{0ti}, A_{0ti}, \bar{F}_{0ti}(t)) \\ &\quad + \sum_t \sum_{i=1}^{n(t)} \sum_{s>t} R(s) D_\beta(s, 0, W_{0ti}, A_{0ti}, \bar{F}_{0ti}(s)). \end{aligned}$$

The time-specific case-control weights are $q_0(t) = P(\tilde{T} = t, \Delta = 1 \mid \tilde{T} \geq t)$ and $\bar{q}_0(t) = 1 - q_0(t)$. If censoring C is independent of T, then $q_0(t) = P(T = t \mid T \geq t)$ is the marginal hazard of T. One can now apply standard maximum likelihood estimation to this log-likelihood, which can be carried out with standard software.

This case-control weighted log-likelihood can also be written down for a semi-parametric Cox-proportional hazards model, and, again, the corresponding maximum likelihood estimator can be found by using standard maximum likelihood estimation software developed for fitting a Cox-model based on prospective sampling.

Finally, in an anologue fashion we can now obtain case-control weighted targeted maximum likelihood estimators of particular parameters of this intensity such as marginal causal effects of A on T, but we reserve this for future work.

Appendix: Tangent space results proving case-control weighted canonical gradient of prospective sampling model equals canonical gradient.

Our results in this section show that the case-control weighted canonical gradient for the prospective sampling model $\mathcal{M}^*$ yields the canonical gradient for the parameter of interest Ψ in the actual case-control sampling model. These results rely on the following assumption. The (typically very large/semiparametric) model $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$ for case-control

design I, and, for case-control design II, $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and m varying over the support of the matching variable M. As a consequence of this canonical gradient representation our proposed case-control weighted targeted maximum likelihood estimator, involving selecting estimators of Q_0^* and g_0^*, under appropriate regularity conditions guaranteeing the wished convergence to a normal limit distribution, is efficient if both of these estimators are consistent, and remains consistent if one of these estimators is consistent.

The results are stated in an incremental fashion thereby building up the proof of the final wished result. As a consequence, most stated results do not require a proof but can be straightforwardly verified.

Tangent space for case-control design I. We start out with presenting the tangent space for case-control design I.

Theorem 8 *(***Tangent space for case-control design I***) Consider case-control design I and the independence model $\mathcal{M}$ described by (3),*

$$dP(P^*)(O) = P^*(W_1, A_1 \mid Y = 1) \prod_j P^*(W_2^j, A_2^j \mid Y = 0),$$

and let $T^*(P^*)$ denote the tangent space at P^* in model $\mathcal{M}^*$. The tangent space at $P(P^*)$ in model $\mathcal{M}$ is given by

$$T_I(P^*) = \\ \left\{ S^*(W_1, A_1, 1) - E^*(S^* \mid Y = 1) + \textstyle\sum_j \{ S^*(W_2^j, A_2^j, 0) - E^*(S^* \mid Y = 0) \} \right\},$$

where S^* varies across $T^*(P^*)$.

Since this tangent space is expressed in terms of the tangent space of the underlying model $\mathcal{M}^*$ we now need to understand the tangent space of $\mathcal{M}^*$. The following theorem fully characterizes this tangent space for models $\mathcal{M}^*$ described by separate models for $P(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$.

Theorem 9 *(***Tangent space for underlying model $\mathcal{M}^*$***) Consider the data structure $O^* = (W, A, Y)$ and model $\mathcal{M}^*$ for its probability distribution. We make the following assumption on $\mathcal{M}^*$: Let $\mathcal{M}^* = \cap_\delta \mathcal{M}^*(\delta)$, where $\mathcal{M}^*(\delta)$ is a model for $P_0^*(W, A \mid Y = \delta)$ indexed by (possibly infinite dimensional) parameter $\theta(\delta)$, for each $\delta \in \{0, 1\}$, and assume that $\theta(\delta)$ for different choices of δ are variation independent parameters.*

DOI: 10.2202/1557-4679.1114

If the marginal distribution $q_0(\delta) = P(Y = \delta)$ of Y is known in model $\mathcal{M}^*$, then, we can represent $T^*(P^*)$ as

$$T^*(P^*) = \sum_\delta T^*_\delta(P^*), \tag{6}$$

where the latter sum-space is an orthogonal sum, and $T^*_\delta(P^*)$ denotes the tangent space generated by $\theta(\delta)$, which can be represented as

$$T^*_\delta(P^*) = \{I(Y = \delta)\,(S^*(W, A, \delta) - E(S^* \mid Y)) : S^* \in T^*(P^*)\}\,.$$

If $q_0(\delta)$ is unknown and modelled, then

$$T^*(P^*) = L^2_0(P^*_Y) \oplus \sum_\delta T^*_\delta(P^*), \tag{7}$$

where $L^2_0(P^*_Y)$ is the Hilbert space of functions of Y with mean zero and finite variance w.r.t. P^*. We also note that for a $S^* \in L^2_0(P^*)$, the projection of S^* on $T^*_\delta(P^*)$ is given by

$$\Pi(S^* \mid T^*_\delta(P^*)) = I(Y = \delta)\,(S^*(W, A, \delta) - E(S^* \mid Y))\,,$$

and the projection of S^* onto $T^*(P^*)$ described by the orthogonal decomposition (7) is given by

$$S^* = E(S^* \mid Y) + \sum_\delta \Pi(S^* \mid T^*_\delta(P^*)).$$

Tangent space for case-control design II. We now present the tangent space for matched case-control design II.

Theorem 10 *(***Tangent space for case-control design II***) Consider case-control design II and the independence model $\mathcal{M}$ described by (4),*

$$dP(P^*)(O) = P^*(M_1)P^*(A_1, W_1 \mid Y = 1, M_1) \prod_j P^*(A^j_2, W^j_2 \mid Y = 0, M_1),$$

and let $T^*(P^*)$ denote the tangent space at P^* in model $\mathcal{M}^*$. The tangent space at $P(P^*)$ in model $\mathcal{M}$ is given by

$$T_{II}(P^*) = L^2_0(M_1) \oplus$$
$$\Big\{S^*(Z_1, 1) - E^*(S^* \mid M = M_1, Y = 1) + \textstyle\sum_j \{S^*(Z^j_2, 0) - E^*(S^* \mid M = M_1, Y = 0)\}\Big\},$$

where S^* varies across $T^*(P^*)$, $Z_1 = (M_1, W_1, A_1)$ and $Z^j_2 = (M_1, W^j_2, A^j_2)$.

Since this tangent space is characterized in terms of the underlying tangent space $T^*(P^*)$ for model $\mathcal{M}^*$ we now fully characterize the latter tangent space for models $\mathcal{M}^*$ described by separate models for $P^*(W, A \mid M = m, Y = \delta)$ for the different values of m and δ.

Theorem 11 *(***Tangent space for model $\mathcal{M}^*$ including matching variable***) We make the following assumption on the model $\mathcal{M}^*$: Suppose that $\mathcal{M}^* = \cap_{m,\delta}\mathcal{M}^*(m,\delta)$, where $\mathcal{M}^*(m,\delta)$ is a model for $P_0^*(W, A \mid M = m, Y = \delta)$ indexed by (e.g., infinite dimensional) parameter $\theta(m,\delta)$, for each $\delta \in \{0,1\}$ and possible outcome m for M, and it is assumed that $\theta(m,\delta)$ are variation independent parameters.*

If $q_0(\delta \mid m) = P(Y = \delta \mid M = m)$ is known and the marginal distribution of M is unspecified in model $\mathcal{M}^*$, then, we can represent $T^*(P^*)$ as

$$T^*(P^*) = L_0^2(M) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{8}$$

where the latter sum-space is an orthogonal sum, and $T^*_{m,\delta}(P^*)$ denotes the tangent space generated by $\theta(m,\delta)$, which can be represented as

$$T^*_{m,\delta}(P^*) = \{I(M = m, Y = \delta)\,(S^*(m, W, A, \delta) - E(S^* \mid M, Y)) : S^* \in T^*(P^*)\}\,.$$

If the conditional distribution $q_0(\delta \mid m)$ of Y, given M, is unknown and modeled, then

$$T^*(P^*) = L_0^2(P^*_M) \oplus T^*(q_0) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{9}$$

where $T^*(q_0)$ denotes the tangent space generated by the scores of the parameters of $q_0(\delta \mid m)$. We also note that for a $S^* \in L_0^2(P^*)$, the projection onto $T^*_{m,\delta}(P^*)$ is given by

$$\Pi(S^* \mid T^*_{m,\delta}(P^*)) = I(M = m, Y = \delta)\,(S^*(m, W, A, \delta) - E(S^* \mid M, Y))\,,$$

and, under the assumption that $q_0(\delta \mid m)$ is unspecified, the projection of S^* onto $T^*(P^*)$ described by the orthogonal decomposition (9) is given by

$$S^* = E(S^* \mid M) + \{E(S^* \mid Y, M) - E(S^* \mid M)\} + \sum_{m,\delta} \Pi(S^* \mid T^*_{m,\delta}(P^*)).$$

DOI: 10.2202/1557-4679.1114

Special score for case-control design I. We will later show that the case-control weighted canonical gradient is in the tangent space $T_I(P^*)$ by selecting a special choice $S^* \in T^*(P^*)$ defined in the next result. The following result shows that this special choice is indeed a member of $T^*(P^*)$.

Result 1 *Let $O^* = (W, A, Y) \sim P_0^* \in \mathcal{M}^*$ and assume that the tangent space $T^*(P^*)$ at $P^* \in \mathcal{M}^*$ is given by orthogonal decomposition (7). Given a $D^* \in T^*(P^*)$, we have*

$$\begin{aligned} S^*(W, A, Y) &= q_0(Y) \{D^*(W, A, Y) - E^*(D^* \mid Y)\} \\ &\in T^*(P^*). \end{aligned}$$

The same applies if $q_0(0)$ is replaced by $q_0(0)/J$.

Proof. Firstly, we note that for each δ, $\Pi(D^* \mid T_\delta(P^*)) \in T^*(P^*)$, and by linearity of the space $T_\delta(P^*)$ (i.e., closure under multiplication by scalar) we have that $q_0(\delta)\Pi(D^* \mid T_\delta^*(P^*)) \in T^*(P^*)$. By linearity of $T^*(P^*)$, it follows thus that

$$\begin{aligned} &\textstyle\sum_\delta q_0(\delta)\Pi(D^* \mid T_\delta^*(P^*)) \\ &= \textstyle\sum_\delta q_0(\delta) I(Y = \delta) \left(D^*(W, A, \delta) - E^*(D^* \mid Y)\right) \\ &= q_0(Y) \left(D^*(W, A, Y) - E^*(D^* \mid Y)\right) \\ &= S^*(W, A, Y) \\ &\in T^*(P^*). \end{aligned}$$

This completes the proof. □

Special score for case-control design II. For case-control design II, we need a similar result.

Result 2 *Consider the model $O^* = (M, W, A, Y) \sim P_0^* \in \mathcal{M}^*$ and let $T^*(P^*)$ denote the tangent space at $P^* \in \mathcal{M}^*$ and assume it satisfies orthogonal decomposition (9). Given a $D^* \in T^*(P^*)$, we have*

$$\begin{aligned} S_m^*(M, W, A, Y) &\equiv I(M = m) q_0(Y \mid m) \{D^*(m, W, A, Y) - E^*(D^* \mid M, Y)\} \\ &\in T^*(P^*). \end{aligned} \tag{10}$$

The same result applies if we replace $q_0(0 \mid m)$ by $q_0(0 \mid m)/J$.

Proof. Firstly, we note that for each m, δ, $\Pi(D^* \mid T_{m,\delta}(P^*)) \in T^*(P^*)$, and by linearity of the space $T_{m,\delta}(P^*)$ (i.e., closure under multiplication by scalar) we have that $q_0(\delta \mid m)\Pi(D^* \mid T^*_{m,\delta}(P^*)) \in T^*(P^*)$. By linearity of $T^*(P^*)$, it follows thus that

$$\begin{aligned}
&\textstyle\sum_\delta q_0(\delta \mid m)\Pi(D^* \mid T^*_{m,\delta}(P^*)) \\
&= \textstyle\sum_\delta q_0(\delta \mid m) I(M = m, Y = \delta)\,(D^*(m, W, A, \delta) - E^*(D^* \mid M, Y)) \\
&= I(M = m) q_0(Y \mid m)\,(D^*(m, W, A, Y) - E^*(D^* \mid M, Y)) \\
&= S^*_m(M, W, A, Y) \\
&\in T^*(P^*).
\end{aligned}$$

This completes the proof. □

Case-control weighted score equals a score, case-control design I. We are now ready to establish our wished results showing that the case-control weighted canonical gradient of the prospective sampling model is an element of the tangent space for the observed data model $\mathcal{M}$.

Theorem 12 *(***Case-control weighted score is a score, Design I***) Consider case-control design I, its independence model $\mathcal{M}$ described by (3), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (7).*

If $D^* \in T^*(P^*)$, then

$$D_{q_0}(O) = q_0 D^*(W_1, A_1, 1) + \frac{(1 - q_0)}{J} \sum_j D^*(W_2^j, A_2^j, 0) \in T_I(P^*).$$

Specifically, if we set

$$S^*(W, A, Y) = q_0(Y)\,\{D^*(W, A, Y) - E^*(D^* \mid Y)\} \in T^*(P^*),$$

where $q_0(Y) = I(Y = 1)q_0 + I(Y = 0)(1 - q_0)/J$, then

$$\begin{aligned}
D_{q_0}(O) \;=\; & S^*(W_1, A_1, 1) - E^*(S^*(W, A, Y) \mid Y = 1) \\
& + \sum_j \{S^*(W_2^j, A_2^j, 0) - E^*(S^*(W, A, Y) \mid Y = 0)\}.
\end{aligned}$$

(Here, we use the fact for $J = 1$, $E^*(S^* \mid Y = 1) + E^*(S^* \mid Y = 0) = 0$.)

This establishes the wished corollary stating that the case-control weighted canonical gradient for the prospective sampling model yields the canonical gradient for the case-control sampling model $\mathcal{M}$.

DOI: 10.2202/1557-4679.1114

Corollary 1 *Consider case-control design I, its independence model $\mathcal{M}$ described by (3), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (7).*

Suppose that $D^*(P^*)$ is the canonical gradient of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d \Psi^*$, and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P(P^*) \in \mathcal{M}$, satisfy $\Psi(P(P^*)) = \Psi^*(P^*)$.

Assume that the corresponding case-control weighted D_{q_0} (satisfies the regularity conditions such that it) is a gradient for Ψ at $P(P^*)$. Then D_{q_0} is the canonical gradient of Ψ at $P(P^*)$.

Case-control weighted score is a score, Case-Control Design II. We establish the same type result for case-control design II.

Theorem 13 *(***Case-control weighted score is a score, Design II***) Consider case-control design II, its independence model $\mathcal{M}$ described by (4), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (9).*

For any $D^* \in L^2(P^*)$, we have

$$\begin{aligned} D_{q_0,\bar{q}_0}(O) &\equiv q_0 D^*(M_1, W_1, A_1, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_j D^*(M_1, W_2^j, A_2^j, 0) \\ &= \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_1 = m) D^*_{m,q_0}, \end{aligned}$$

where

$$D^*_{m,q_0}(O) \equiv q_0(1 \mid m) D^*(m, W_1, A_1, 1) + \frac{q_0(0 \mid m)}{J} D^*(m, W_2^j, A_2^j, 0).$$

For each m, and $D^* \in T^*(P^*)$, we have

$$I(M_1 = m) D^*_{m,q_0} \in T_{II}(P^*)$$

so that it follows that

$$D_{q_0,\bar{q}_0}(P^*) \in T_{II}(P^*).$$

Let $q_{0J}(\delta \mid m) = q_0(1 \mid m)\delta + (1-\delta)q_0(0 \mid m)/J$. Specifically, if we set

$$S^*_m(M, W, A, Y) = I(M = m) q_{0J}(Y \mid m) \left\{ D^*(m, W, A, Y) - E^*(D^* \mid M, Y) \right\},$$

which is an element of $T^*(P^*)$ by (10) above, then

$$\begin{aligned} I(M_1 = m) D^*_{m,q_0}(O) &= S^*_m(M_1, W_1, A_1, 1) - E^*(S^*_m \mid M, Y = 1) \\ &\quad + \sum_j \{ S^*_m(M_1, W_2^j, A_2^j, 0) - E(S^* \mid M, Y = 0) \} \\ &\in T_{II}(P^*). \end{aligned}$$

Here we use that for any $D^* \in L_0^2(P^*)$,

$$q_0(1 \mid m)E^*(D^* \mid M = m, Y = 1) + q_0(0 \mid m)E(D^* \mid M = m, Y = 0) = 0.$$

This gives us the wished result.

Corollary 2 *(***Case-control weighted canonical gradient is a canonical gradient, Design II***) Consider case-control design II, its independence model* $\mathcal{M}$ *described by (4), and assume the tangent space* $T^*(P^*)$ *of* $\mathcal{M}^*$ *at* P^* *satisfies the orthogonal decomposition (9).*

If $D^*(P^*)$ *is the canonical gradient of* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ *at* P^**, then*

$$\begin{aligned} D_{q_0,\bar{q}_0} &\equiv \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_1 = m) D^*_{m,q_0} \\ &\in T_{II}(P^*). \end{aligned}$$

Thus, under the conditions for which which $D_{q_0,\bar{q}_0}(P^*)$ *is a gradient of* $\Psi : \mathcal{M} \to \mathbb{R}^d$ *at* $P(P^*) \in \mathcal{M}$*, satisfying* $\Psi(P(P^*)) = \Psi^*(P^*)$ *for specified parameter* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$*, we also have that* $D_{q_0,\bar{q}_0}(P^*)$ *is the canonical gradient of* Ψ *at* $P(P^*)$.

References

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

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

P.J. Bickel, C.A. J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, MD, 1993. ISBN 0-8018-4541-6.

N.E. Breslow. Statistics in epidemiology: the case-control study. *J Am Stat Soc*, 91:14–28, 1996.

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies*. International Agency for Research on Cancer, Lyon, 1980.

N.E. Breslow, N.E. Day, K.T. Halvorsen, R.L. Prentice, and C. Sabal. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epid*, 108(4):299–307, 1978.

N.E. Breslow, J.M. Robins, and J.A. Wellner. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6(3): 447–455, 2000.

D. Collett. *Modeling Binary Data*. Chapman and Hall, London, 1991.

J. Cornfield. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *J Nat Cancer Inst*, 11:1269–1275, 1951.

J. Cornfield. A statistical problem arising from retrospective studies. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium, Volume IV*, pages 133–148. University of California Press, 1956.

S.R. Cosslett. Maximum likelihood estimator for choice-based samples. *Econometrica*, 49(5):1289–1316, 1981.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

T.R. Holford, C. White, and J.L. Kelsey. Multivariate analysis for matched case-control studies. *Am J Epid*, 107(3):245–255, 1978.

D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, New York, 2nd edition, 2000.

N.P. Jewell. *Statistics for Epidemiology*. Chapman and Hall/CRC, Boca Raton, 2004.

C.F. Manski and S.R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988, 1977.

C.F. Manski and D. McFadden. Alternative estimators and sample designs for discrete choice analysis. In C.F. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*. The MIT Press, Cambridge, MA, 1981.

R. Mansson, M.M. Joffe, W. Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and cohort studies. *American Journal of Epidemiology*, 00:1–8, 2007.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. Technical report 215, Division of Biostatistics, University of California, Berkeley, April 2007.

A.P. Morise, G.A. Diamon, R. Detrano, M. Bobbio, and Erdogan Gunel. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

S. Newman. Causal analysis of case-control data. *Epid Persp Innov*, 3:2, 2006. URL `http://www.epi-perspectives.com/content/3/1/2`.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994.

K. Rothman and S. Greenland. *Modern Epidemiology.* Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

D.B. Rubin. *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 2006.

J.J. Schlesselman. *Case-Control Studies: Design, Conduct, Analysis.* Oxford University Press, Oxford, 1982.

M.J. van der Laan. Causal effect models for intention to treat and realistic individualized treatment rules. Technical report 203, Division of Biostatistics, University of California, Berkeley, 2006.

M.J. van der Laan and J.M. Robins. Unified methods for censored longitudinal data and causality. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

S. Wachholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.