DOI: 10.1111/biom.12222

Case-Base Methods for Studying Vaccination Safety

Olli Saarela[1,*] and James A. Hanley[2]

[1]Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th floor, Toronto, Ontario, Canada M5T 3M7

[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Avenue West, Montreal, Quebec, Canada H3A 1A2

[*]*email:* olli.saarela@utoronto.ca

Summary. Pooling of controls under nested-case control settings can produce substantial efficiency gains compared to standard time-matched analysis using the Mantel–Haenszel method or conditional logistic regression. In the context of possible adverse effects of early childhood vaccinations, we propose pooling of the information from the controls to estimate the population exposure prevalence as a parametric or nonparametric function of time, and possibly other factors. This function in turn may be used as a plug-in estimate to control for confounding in the subsequent estimation of rate ratios. We derive standard errors for the resulting two-step estimators, demonstrate through simulations the efficiency gains compared to standard matched analysis, and propose a novel graphical presentation of the vaccination and adverse event time data. We formulate the methods in the general framework of case-base sampling, which subsumes the different case-control and case-only methods.

Key words: Case-base study; Efficiency; Etiologic study; Nested case-control study; Self-matching; Vaccination study; Variance estimation.

1. Introduction

1.1. *Background*

The study carried out in Mexico and Brazil on the role of rotavirus vaccination in the occurrence of intussusception (Patel et al., 2011) is a good example of the classic time-matched case-control study of adverse effects of vaccinations. Four controls per case, matched by age and other factors, were selected, and classified as exposed in the 7 days following vaccination and unexposed otherwise. Customary matched analysis of such data may be carried out by using Mantel and Haenszel (1959) or conditional logistic regression (e.g., Langholz and Goldstein, 1996) methods.

Pooling of controls under time-matched sampling (a.k.a. nested case-control study, risk set sampling, or incidence density sampling), originally suggested by Samuelsen (1997), has been extensively discussed in the literature; reviews of various methods are contained for instance in Breslow and Wellner (2007), Samuelsen, Ånestad, and Skrondal (2007), Saarela et al. (2008), and Gray (2009). What is common to these methods is the existence of an enumerable study population or sampling frame, for instance in the form of a cohort, onto which the inferences can be generalized using either weighting by the inverses of the inclusion probabilities or likelihood-based missing data methods.

Compared to standard analysis through conditional logistic regression, where the riskset at each event time involves only the sampled controls, as well as the case itself, pooling of the controls does result in improved efficiency. This is also our motivation for pursuing alternatives to standard matched analyses. However, instead of weighting, our approach is based on estimating the population exposure prevalence, which does not require an enumerable sampling frame or determination of the inclusion probabilities. Our approach is valid irrespective of the particular sampling scheme, as long as the sampling is independent of the exposure status and the eventual event status. This is true for instance under case-cohort (Prentice, 1986) sampling schemes, and under nested case-control sampling schemes if the outcome event does not terminate the follow-up (Section 2.2.2).

Our approach based on estimation of the population exposure prevalence function also enables a novel graphical presentation for the vaccination and adverse event time data. Conventional one-dimensional graphical presentations of temporal association between the vaccination time and event time distributions, such as Figure 1 in the study of H1N1 vaccination and childhood narcolepsy reported by Nohynek et al. (2012), do not enable direct comparison of incidence density in the exposed versus unexposed population time. For this purpose, we use a two-dimensional population-time plot, which is free of confounding due to age, the main time scale in our analysis. This is especially important in studies of scheduled early childhood vaccinations, where age is a major confounder. A conventional graphical display for temporal association is obtained when the two-dimensional presentation is collapsed over the population dimension.

1.2. *Concepts, Notation, and Plan of the Article*

The *etiologic study* characterized by Miettinen and Karp (2012) can be seen as a unified framework for understanding the sampling aspect of epidemiological study designs. To place their concepts into an appropriate mathematical context, we first define a history of follow-up experience as a generated σ-algebra $\mathcal{F}_{it} = \sigma\{N_i(u), Z_i(u), X_i : 0 \leq u \leq t\}$ (e.g., Aalen, Borgan, and Gjessing, 2008, p. 43), where $N_i(t)$ is a counting process for the incident adverse events, $Z_i(t)$ is the exposure status at time t (say, a vaccination within 1 week before t), and x_i is a vector of potential confounders other than age, for example, gender, socioeconomic status, and other demographic factors. Henceforth, such a history at a "person-coordinate" i and "time-coordinate" t is referred to as a *person-moment* (cf. Miettinen and Karp, 2012, 94–95). Aggregation of person-moments over time t and individuals i will be referred to as *population-time*. The *study base* comprises the aggregate population-time contributed by all individuals in the study population of size N over the study period $(0, \tau]$, that is, $\bigcup_{i=1}^{N} \mathcal{F}_{i\tau}$. The *case series* is the discrete set of person-moments where $\mathrm{d}N_i(t) = 1$. To enable comparisons, a random sample of *base series* person-moments is drawn randomly from the study base.

Drawing the base series independently of the case series preserves the connection between the study and its base, enabling a wider variety of possible analysis methods. One such alternative is pooling the information from the base series to estimate the population exposure prevalence as a parametric or non-parametric function of time, and possibly other factors, as noted in Section 1.1. This function in turn may be used as a plug-in estimate to control for confounding in the subsequent estimation of rate ratios. To define the parameter of interest, for children t days of age, we denote the adverse event incidence rates in the index and reference categories of exposure by $\lambda_1(t, x_i)$ and $\lambda_0(t, x_i)$, given by $\lambda_j(t, x_i) \equiv P(\mathrm{d}N_i(t) = 1 \mid N_i(t-), Z_i(t) = j, x_i)/\mathrm{d}t,\ j \in \{0, 1\}$. The object of inference, the rate ratio $\theta \equiv \lambda_1(t, x_i)/\lambda_0(t, x_i)$, is assumed proportional across the values of t and x_i. For notational simplicity, we assume censoring to be of type I, due to the end of the follow-up period at time τ. Age is used as the time scale of the analysis throughout.

The study reported in Patel et al. (2011) employed both case-control and self-matched approaches and found the results to be in agreement. However, the self-matched methods have the advantage of model-free controlling for time-invariant confounders, and although less suitable for studies of chronic disease outcomes, they are often preferred in vaccination studies, where the exposures are transient and outcomes often recurrent (Farrington, 1995; Whitaker, Hocine, and Farrington, 2009). For this reason, we also outline a self-matched case-base sampling approach, where the base series is drawn as a sample of person-moments from the follow-up time contributed by the individuals with an adverse outcome event.

In Section 2.1, we describe the estimation of θ in the situation where the vaccination histories are *readily available on the whole study population*; in Section 2.2, we address the situation where the population exposure prevalence is unknown and needs to be *estimated using a base series sample*. In Sections 2.3 and 2.4 we outline an estimation procedure that aims to extract more information from the base series, and in Section 2.5 we extend this to deal with matching factors other than time. In Section 3, we propose a self-matched version of the case-base sampling method of Hanley and Miettinen (2009). This is followed by a simulation study on the efficiency gains in Section 4 and a discussion in Section 5.

2. Rate Ratio Estimation

2.1. *When Vaccination Histories Are Readily Available on the Whole Study Population*

Some countries such as Finland (e.g., Nohynek et al., 2012) maintain computerized records that allow researchers to readily assemble the vaccination histories of all children in the study population, and consequently, the population exposure prevalence at any given age t. The population exposure prevalence at age t for children with the covariate profile x_i is defined as $\pi(t, x_i) \equiv P(Z_i(t) = 1 \mid X_i = x_i)$. Suppose that the observed information on n incident cases of disease observed during the study period consists of the ordered event times $t_1, \ldots, t_n$ and the corresponding exposure status $Z_i(t_i)$, $i = 1, \ldots, n$. Further, assume that the events are generated by a non-homogeneous Poisson process with rate $\lambda_{Z_i(t)}(t, x_i)$, so that the rate is not modified by the past history of the process. Conditioning on an event having occurred at time t_i and the covariate profile x_i results in conditional likelihood contributions of the form

$$\begin{aligned} P(Z_i(t_i) = 1 \mid \mathrm{d}N_i(t_i) = 1, x_i) &= \frac{\lambda_1(t, x_i) P(Z_i(t_i) = 1 \mid x_i)}{\sum_{j=0}^{1} \lambda_j(t, x_i) P(Z_i(t_i) = j \mid x_i)} \\ &= \frac{\theta\pi(t_i, x_i)}{1 - \pi(t_i, x_i) + \theta\pi(t_i, x_i)} \\ &\equiv \mu(t, x_i), \end{aligned}$$

so that $Z_i(t) \mid (\mathrm{d}N_i(t) = 1, x_i) \sim \text{Bernoulli}(\mu(t, x_i))$. If the events are not generated by a Poisson process, for instance when the follow-up is terminated by the first incident event, this likelihood still applies under the null $\theta = 1$, and approximately when the outcome event is rare, or when the exposure is transient, so that it does not substantially alter the event-free survival probability. If the exposure prevalences $\pi(t, x_i)$, $i = 1, \ldots, n$, are known, an estimate for θ may be obtained through maximization of the conditional likelihood expression

$$L(\theta; \pi) \overset{\theta}{\propto} \prod_{i=1}^{n} \frac{[\theta\pi(t_i, x_i)]^{Z_i(t_i)}}{1 - \pi(t_i, x_i) + \theta\pi(t_i, x_i)}. \tag{1}$$

Likelihood expressions such as (1) have been called "case-pseudocontrol" or "case-distribution" likelihoods by Greenland (1999), but are conditional likelihoods in the general meaning of the term (as defined by Cox and Hinkley, 1974, pp. 16–17). Another connection worth pointing out is that $\pi(t, x_i)$ is a continuous time analogue of the propensity score (Rosenbaum and Rubin, 1983), and appears naturally in conditional likelihoods of the type (1) without the need for special propensity score adjustment methods, for example, through regression adjustment, matching or stratification (cf. Månsson et al., 2007).

The maximization of (1) can be carried out using standard conditional logistic regression software that can accommodate weights or offset terms. Alternatively, since $\text{logit}\{\mu(t, x_i)\} = \text{logit}\{\pi(t, x_i)\} + \log(\theta)$, the parameter θ can be fitted using unconditional logistic regression with only an intercept term, with $Z_i(t)$ as the outcome variable and $\text{logit}\{\pi(t, x_i)\}$ as the offset term. The fitted intercept, when exponentiated, serves as the conditional ML estimate of θ, and the precision of this can be measured as usual by inverting the observed information at the maximum likelihood point. Yet another alternative would be the continuous-time counterpart of the familiar Mantel and Haenszel (1959) estimator with known denominators, of the form

$$\hat{\theta}_{\mathrm{MH}} = \frac{\sum_{i=1}^{n} Z_i(t_i)[1 - \pi(t_i, x_i)]}{\sum_{i=1}^{n}[1 - Z_i(t_i)]\pi(t_i, x_i)}. \tag{2}$$

We note that the conditional likelihood (1) and the estimator (2) feature contributions only from the children with at least one adverse event during the study period, and thus, taking $\pi(t, x_i)$ to be known, these approaches may be characterized as "case-only" or "case series." In the next subsection we will address the situation where $\pi(t, x_i)$ is unknown and needs to be estimated using data on a base series sample. (Indeed, this can be seen as the very purpose for which the base series is required to complement the case series.) Before this, to see how much information is lost compared to the setting with $\pi(t, x_i)$ known, we note (see Supplementary Appendix A) that the observed information on the log-rate ratio $\eta = \log(\theta)$ based on the conditional likelihood (1) is given by

$$I^{\eta\eta} = \sum_{i=1}^{n} \frac{\exp(\eta)\pi(t_i, x_i)[1 - \pi(t_i, x_i)]}{[1 + (\exp(\eta) - 1)\pi(t_i, x_i)]^2}.$$

As an example, we use the vaccination pattern shown in Figure 1, given by a gamma distribution with shape = 5 and scale = 0.5, with the eventual proportion of vaccinated being 75%, and exposure defined as a vaccination within the previous week. We use a rescaled gamma distribution with shape = 20 and scale = 4.5 as the baseline hazard function to mimic a temporal distribution of 300 background cases in a population of $N = 10{,}000$ individuals followed up for 20 weeks, a distribution with a slightly longer right tail and a mean located at roughly 90 days. Taking the log rate ratio to be $\eta = 1.5$, we simulate 374 cases in all, giving $I_\eta = 54.4$, $\hat{V}(\hat{\eta}) = 1/54.4$, and a 95% multiplicative margin of error for $\hat{\theta}$ of

$$\mathrm{MME} \equiv \exp\left\{z_{\alpha/2}\sqrt{\hat{V}(\hat{\eta})}\right\} - 1 = \exp\{1.96\sqrt{1/54.4}\} - 1,$$

or 30%.

Figure 1 shows how the full information on both the study base and the case series can be presented graphically in a population-time plot. The 100% × 20 weeks rectangle in the upper panel represents the total population-time comprising the study base, which is further split into exposed and unexposed population-time. Comparison of the daily exposure prevalence in the middle panel and the daily incident cases in the bottom panel represents the conventional visual presentation for the temporal association between the exposure and the incidence, whereas the proposed population-time presentation in the top panel enables direct comparison of incidence density in the exposed and unexposed population-time, free of confounding by age.

2.2. *When it Requires Substantial Effort to Assemble the Vaccination Histories*

2.2.1. Traditional nested case-control study. The case series remains unchanged, but the unknown exposure prevalences $\pi(t_i, x_i)$ at the event times t_i, $i = 1, \ldots, n$, must now be estimated based on information provided by a base series of person-moments drawn from the (possibly stratified by x_i) risk sets at the event times. A sample of size m is drawn independently at each event time $t_1, \ldots, t_n$, and irrespective of the eventual event status of the at-risk individuals, from the stratified risk set, and classified into $H_1(t_i)$ exposed and $H_0(t_i)$ unexposed person-moments. The exposure prevalence function at each event time may now be estimated by $\hat{\pi}(t, x_i) = H_1(t_i)/m$, or equivalently, by $\mathrm{logit}\{\hat{\pi}(t, x_i)\} = \log\{H_1(t_i)/H_0(t_i)\}$. The M–H estimate for the rate ratio may now be computed by substituting in (2) $H_1(t_i)$ and $H_0(t_i)$ for $\pi(t, x_i)$ and $1 - \pi(t, x_i)$, respectively, combined with either the test-based (Miettinen, 1976) or Robins–Breslow–Greenland (RBG) variance estimator (Robins, Breslow, and Greenland, 1986; see also Silcocks, 2005). Alternatively, conditional maximum likelihood estimates may be obtained through conditional logistic regression. In either case, the standard matched analysis does not use optimally the available information on the base series, since using only the individually matched person-moments (say, $m = 4$ of them at each event time) for the estimation of the exposure prevalences makes these inputs, and as a result, the rate ratio estimator, highly variable, as will be demonstrated in Section 4.

2.2.2. Extracting more information from the base series. As implied above, substantial efficiency gains are possible if the vaccination time data on the whole base series can be pooled in the estimation of the exposure prevalence function $\pi(t, x_i)$. This is especially the case in the present context, where the childhood vaccinations may be expected to follow a smooth pattern over age; soon after the birth of the infant, maternal and child health personnel, using the recommended vaccinations schedules, set up appointments for parents to have their infants vaccinated. Thus, even if individual compliance deviates slightly from this because of sickness and other unexpected events, the generally smooth distribution of the children's ages at the time of vaccination generates a generally smooth exposure pattern, which enables efficiency gains when pooling the base series information. Smoothness is not a required assumption in the semi-parametric two-step approach introduced in Section 2.4, but the efficiency gain naturally depends on the degree of smoothness.

Let V stand for the observed vaccination time of an individual in the study population, with an indicator variable A recording whether the individual was eventually vaccinated or not. For simplicity we assume that the end of the study period is after the scheduled vaccination period, so that $\{A = 1\} \equiv \{V \in [0, \tau)\}$. Finally, let R be the inclusion indicator for this individual contributing at least one person-moment to the sampled base series. Since only part of the population will eventually be vaccinated, we opt to model the vaccination time through the conditional distribution $P(V \in \mathrm{d}v \mid A = 1, x)$ and the marginal probability $P(A = 1)$. Another alternative, which we do not pursue here, would be to model directly the improper vaccination time

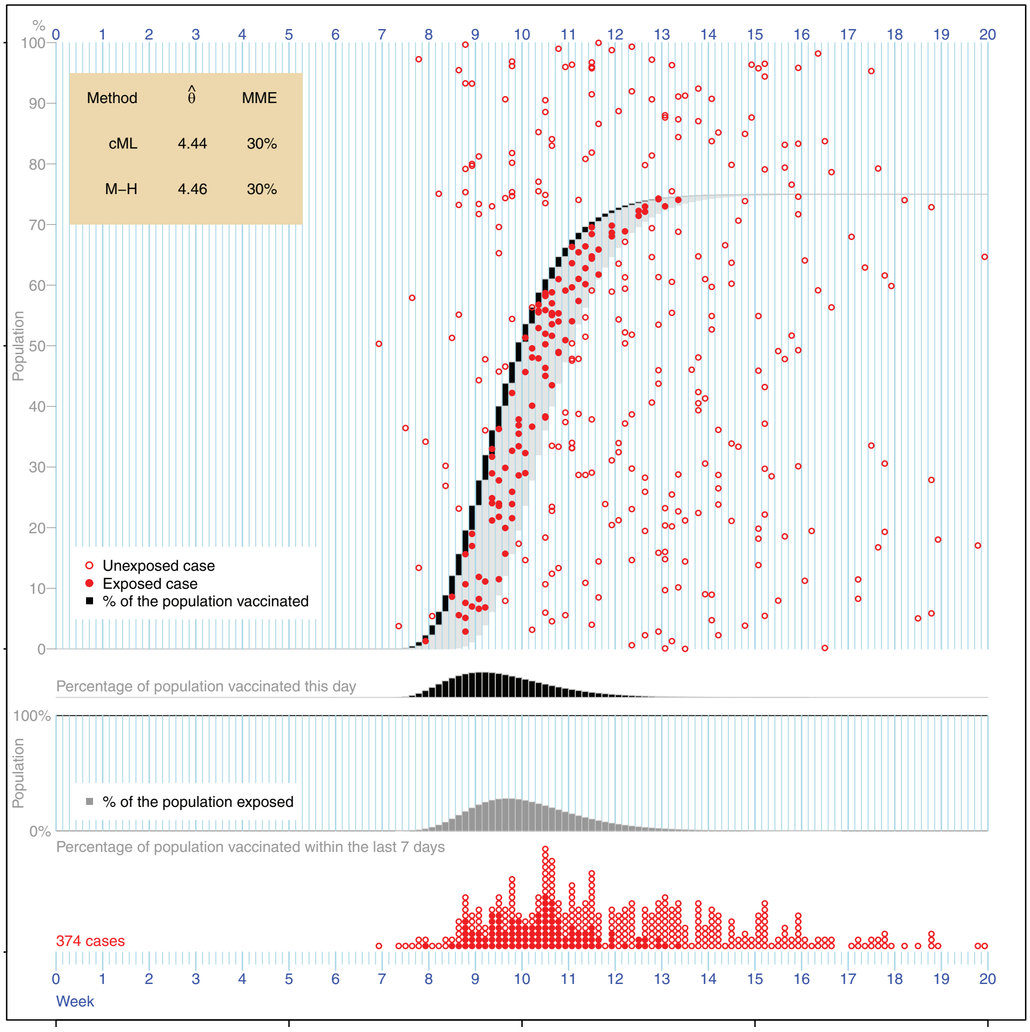


Figure 1. **Inset:** conditional Maximum Likelihood (cML) and Mantel–Haenszel (M–H) rate ratio estimates and their associated multiplicative margins of error. **Top panel:** Population-time plot showing the unexposed (unshaded) and exposed (shaded) population-time comprising the study base. The bars represent the proportion of the population vaccinated on each day, and thus remaining exposed for the following week. In addition, shown are exposed (filled circles) and unexposed (unfilled circles) cases that occurred in each day, ordered on the y-axis by their vaccination times relative to the population vaccination distribution. (The eventually unvaccinated cases are spread uniformly over the y-axis from 75% to 100%.) **Middle panel:** The bars show the proportion of the population exposed at the midpoint of each day (vaccinated within the previous week). **Bottom panel:** The counts of exposed and unexposed cases collapsed over the population dimension to the time dimension. This figure appears in color in the electronic version of this article.

distribution $P(V \in \mathrm{d}v \mid x)$, the connection between the two approaches being $P(V \in \mathrm{d}v \mid A = 1, x) = P(V \in \mathrm{d}v \mid x)/P(V \in [0, \tau) \mid x)$.

Pooling of the vaccination data is valid irrespective of the specific sampling mechanism used in the sampling of the base series, as long as the mechanism is such that $R \perp\!\!\!\perp (V, A) \mid x$. To see this, we note that the conditional likelihood to be used for estimation of the vaccination time distribution, which determines the exposure prevalence function, is then $P(V \in \mathrm{d}v, A \mid R = 1, x) = P(V \in \mathrm{d}v, A \mid x)$, the population vaccination distribution. This conditional independence also implies that the selection mechanism must be independent of the history $\sigma\{N(t) : 0 \leq t \leq \tau\}$ of the outcome process whenever the outcome is not independent of the exposure. This is true under unstratified and stratified (by x) case-cohort sampling schemes, and under nested case-control sampling schemes, whenever the outcome event does not terminate the follow-up (which is the case under the Poisson process).

In contrast, in a case-control study the controls would be chosen from the event-free individuals (with $N(\tau) = 0$). If it can be assumed that the adverse event is rare in the sense that $P(V \in \mathrm{d}v, A \mid N(\tau) = 0, x) \approx P(V \in \mathrm{d}v, A \mid x)$, pooling of the controls is still valid approximately, if the selection mechanism is such that $R \perp\!\!\!\perp (V, A) \mid (N(\tau) = 0, x)$. This follows because now $P(V \in \mathrm{d}v, A \mid R = 1, N(\tau) = 0, x) = P(V \in \mathrm{d}v, A \mid N(\tau) = 0, x) \approx P(V \in \mathrm{d}v, A \mid x)$.

With these assumptions on the sampling mechanism, the exposure prevalence functions in (2) may now be replaced with the plug-in estimates $\hat{\pi}(t, x_i)$ drawn from the pooled base series data. However, the usual variance estimators based on the inverse of the observed information will no longer be valid, since they do not account for the estimation cost of the plug-in estimates. Thus, in the following we describe a valid method to approximate the variance of the resulting rate ratio estimator, and compare the efficiency of the alternative methods in a simulated setting.

2.3. *Parametric Two-Step Estimation*

Consider the nested case-control setting where $j = 1, \ldots, m_i$ controls are time-matched to case i at time t_i (usually $m_i = m$ is constant). The observed vaccination times for the matched controls are $(V_{i1}, \ldots, V_{im_i})$, with $(A_{i1}, \ldots, A_{im_i})$ indicating the eventual vaccination status. For the time being, we assume that there are no further matching factors in addition to age, the generalizations allowing this to be considered separately in Section 2.5.

The population exposure prevalence, the proportion of children of age t who have been vaccinated during the past week, is now given by $\pi(t) = P(A = 1) \int_{v \in (t-7,t]} P(V \in \mathrm{d}v \mid A = 1) = \alpha F(t) - \alpha F(t-7)$. Thus, the exposure prevalence is a deterministic function of the distribution function F for vaccination time and the eventual vaccination prevalence α. For the sake of discussion we can parametrize the vaccination time distribution as $P(V \in \mathrm{d}v \mid A = 1; \gamma) P(A = 1; \alpha)$, in which case the exposure prevalence $\pi(t; \gamma, \alpha)$ is also a (deterministic) function of γ and α. Denoting

$$L(\gamma) \equiv \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(V_{ij} \in \mathrm{d}v_{ij} \mid A_{ij} = 1; \gamma)^{A_{ij}}$$

and

$$L(\alpha) \equiv \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(A_{ij}; \alpha),$$

these parameters may be estimated by maximum likelihood as $\hat{\gamma} \equiv \arg\max_\gamma L(\gamma)$ and $\hat{\alpha} \equiv \arg\max_\alpha L(\alpha)$. Using these as plug-in estimates, the log-rate ratio parameter can then be estimated using the conditional likelihood (1) as $\hat{\eta}(\hat{\gamma}, \hat{\alpha}) \equiv \arg\max_\eta L(\eta; \hat{\gamma}, \hat{\alpha})$, with $\pi(t) = \pi(t; \hat{\gamma}, \hat{\alpha})$. In the following, we assume that the overlap of individuals between the sampled risksets is negligible, so that they can be assumed independent. (This is reasonable in the vaccination safety context where the size of the study population N is usually much larger than the number of cases n). In the Supplementary Appendix A we show that through an M-estimator Taylor expansion (e.g., Stefanski and Boos, 2002) around the true values $(\eta_0, \gamma_0, \alpha_0)$, with the number of cases $n \to \infty$ we have that

$$\begin{aligned}\sqrt{n}(\hat{\eta}(\hat{\gamma}, \hat{\alpha}) - \eta_0) &\xrightarrow{\mathrm{d}} N(0,\ E[-I_i^{\eta\eta}(\eta_0; \gamma_0, \alpha_0)]^{-1} \\ &\quad \times V[B_i(\eta_0, \gamma_0, \alpha_0)] E[-I_i^{\eta\eta}(\eta_0; \gamma_0, \alpha_0)]^{-1}),\end{aligned}$$

where

$$\begin{aligned}&V[B_i(\eta_0, \gamma_0, \alpha_0)] \\ &\quad = E[-I_i^{\eta\eta}(\eta_0; \gamma_0, \alpha_0)] \\ &\qquad + E[I_i^{\eta\gamma}(\eta_0; \gamma_0, \alpha_0)] E[-I_i^{\gamma\gamma}(\gamma_0)]^{-1} E[I_i^{\eta\gamma}(\eta_0; \gamma_0, \alpha_0)'] \\ &\qquad + E[I_i^{\eta\alpha}(\eta_0; \gamma_0, \alpha_0)] E[-I_i^{\alpha\alpha}(\alpha_0)]^{-1} E[I_i^{\eta\alpha}(\eta_0; \gamma_0, \alpha_0)']\end{aligned}$$

and the notations are

$$\begin{aligned}I^{\eta\eta}(\eta; \gamma, \alpha) &\equiv \partial^2 \log L(\eta; \gamma, \alpha)/\partial\eta^2, \\ I^{\eta\gamma}(\eta; \gamma, \alpha) &\equiv \partial^2 \log L(\eta; \gamma, \alpha)/\partial\eta\partial\gamma, \\ I^{\eta\alpha}(\eta; \gamma, \alpha) &\equiv \partial^2 \log L(\eta; \gamma, \alpha)/\partial\eta\partial\alpha, \\ I^{\gamma\gamma}(\gamma) &\equiv \partial^2 \log L(\gamma)/\partial\gamma^2 \\ I^{\alpha\alpha}(\alpha) &\equiv \partial^2 \log L(\alpha)/\partial\alpha^2.\end{aligned}$$

This motivates the variance estimator

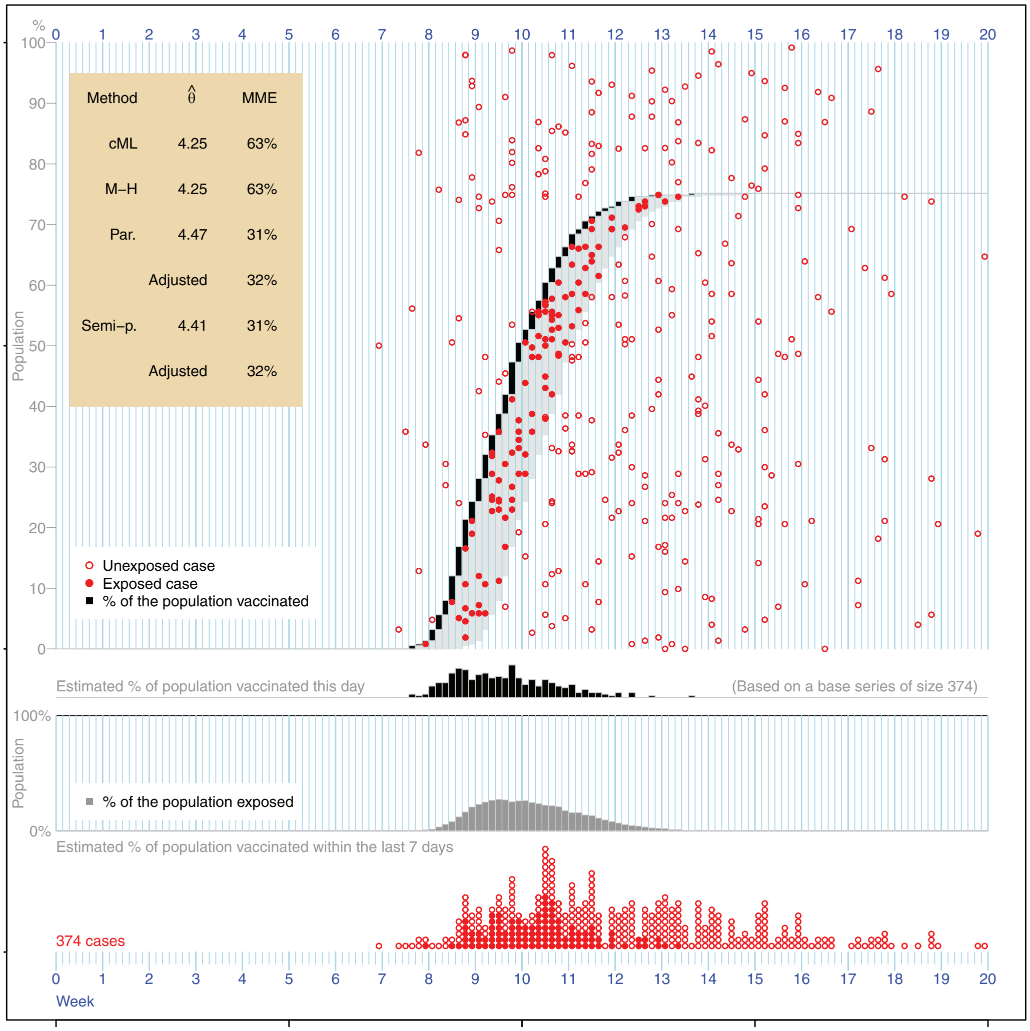
$$\begin{aligned}\hat{V}[\hat{\eta}] = &-I^{\eta\eta}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})^{-1} \\ &+ I^{\eta\eta}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})^{-1} I^{\eta\gamma}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})[-I^{\gamma\gamma}(\hat{\gamma})]^{-1} I^{\eta\gamma}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})' \\ &\times I^{\eta\eta}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})^{-1} \\ &+ I^{\eta\eta}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})^{-1} I^{\eta\alpha}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})[-I^{\alpha\alpha}(\hat{\alpha})]^{-1} I^{\eta\alpha}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})' \\ &\times I^{\eta\eta}(\hat{\eta}; \hat{\gamma}, \hat{\alpha})^{-1}, \qquad (3)\end{aligned}$$

where the first term is the unadjusted variance, and the last two terms penalize for the estimation of the nuisance parameters γ and α.

2.4. *Semi-Parametric Two-Step Estimation*

Since we do not actually wish to assume a parametric distribution for the vaccination time distribution F, this may be estimated by the empirical cumulative distribution function (ECDF) $\hat{F}(t) \equiv \frac{1}{\hat{\alpha}M} \sum_{i=1}^{n} \sum_{j=1}^{m_i} A_{ij} \mathbf{1}_{\{V_{ij} < t\}}$, where $M = \sum_{i=1}^{n} m_i$ is the total number of sampled controls. In the presence of censored vaccination times, the Kaplan–Meier estimator may be used instead. Furthermore, the eventual proportion of vaccinated may be estimated by $\hat{\alpha} = \frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{m_i} A_{ij}$. The estimate for η is again obtained by using the estimated exposure prevalences $\pi(t; \hat{F}, \hat{\alpha})$ as plug-in estimates in (1). We know that here $\sqrt{\hat{\alpha}M}(\hat{F}(t) - F(t)) \xrightarrow{\mathrm{d}} N(0, F(t)(1 - F(t)))$ at any given point t (van der Vaart, 1998, p. 265). The function F is now an infinite dimensional nuisance parameter, but since the conditional likelihood needs to be evaluated only at finite number of timepoints, we can treat the nuisance function as if it was finite dimensional (cf. Farrington and Whitaker, 2006, p. 564), and proceed with a Taylor expansion analogous to the previous section. Using the notations

$$\begin{aligned}U^{\eta}(\eta; F, \alpha) &\equiv \partial \log L(\eta; F, \alpha)/\partial\eta, \\ I^{\eta\eta}(\eta; F, \alpha) &\equiv \partial^2 \log L(\eta; F, \alpha)/\partial\eta^2, \\ I^{\eta F}(\eta; F, \alpha) &\equiv \partial^2 \log L(\eta; F, \alpha)/\partial\eta\partial F,\end{aligned}$$



Method	$\hat{\theta}$	MME
cML	4.25	63%
M–H	4.25	63%
Par.	4.47	31%
	Adjusted	32%
Semi–p.	4.41	31%
	Adjusted	32%

Figure 2. **Inset:** conditional Maximum Likelihood (cML), Mantel–Haenszel (M–H), and parametric (Par.) and semi-parametric (Semi-p.) two-step estimates for the rate ratio and their associated multiplicative margins of error. **Top panel:** Population-time plot showing the unexposed (unshaded) and exposed (shaded) population-time comprising the study base. The bars represent the proportion of the population *estimated* to be vaccinated on each day based on a base series sample of size 374. In addition, shown are exposed (filled circles) and unexposed (unfilled circles) cases that occurred in each day, ordered on the y-axis by their vaccination times relative to the estimated population vaccination distribution. The eventually unvaccinated cases are spread uniformly over the y-axis from 75%, the estimated eventual proportion of vaccinated, to 100%. See Figure 1 for explanation of the middle and bottom panels. This figure appears in color in the electronic version of this article.

where the last derivative is to be understood as pointwise differentiation of $U^{\eta}(\eta; F, \alpha)$ with respect to $F(t)$ at finitely many points t, we arrive at the variance estimator

$$\begin{aligned}\hat{V}[\hat{\eta}] = &-I^{\eta\eta}(\hat{\eta}; \hat{F}, \hat{\alpha})^{-1}\\ &+ I^{\eta\eta}(\hat{\eta}; \hat{F}, \hat{\alpha})^{-1} I^{\eta F}(\hat{\eta}; \hat{F}, \hat{\alpha}) V[\hat{F}] I^{\eta F}(\hat{\eta}; \hat{F}, \hat{\alpha})'\\ &\times I^{\eta\eta}(\hat{\eta}; \hat{F}, \hat{\alpha})^{-1}\\ &+ I^{\eta\eta}(\hat{\eta}; \hat{F}, \hat{\alpha})^{-1} I^{\eta\alpha}(\hat{\eta}; \hat{F}, \hat{\alpha}) V[\hat{\alpha}] I^{\eta\alpha}(\hat{\eta}; \hat{F}, \hat{\alpha})'\\ &\times I^{\eta\eta}(\hat{\eta}; \hat{F}, \hat{\alpha})^{-1}\end{aligned}$$

(see Supplementary Appendix A). Here the covariance terms in the variance–covariance matrix $V[\hat{F}]$ are given by $\frac{1}{\hat{\alpha}M}[F(t_i \wedge t_j) - F(t_i)F(t_j)]$ (van der Vaart, 1998, p. 266), and would in practice be estimated by substituting in the estimated $\hat{F}$ values. In addition, $V[\hat{\alpha}] \approx \hat{\alpha}(1-\hat{\alpha})/M$. We note that the variance estimator obtained in the semi-parametric case is the direct analogue of (3). The partial derivatives required for evaluation of this are found analytically and are given in Supplementary Appendix A. Importantly, we note also that the use of this variance expression does not require inverting an information matrix involving partial derivatives with respect to the high-dimensional nuisance parameter.

Figure 2 shows the same case series as Figure 1, but replaces the full vaccination information on the study base by the vaccination information on a base series of size 374. The inset shows the traditional conditional Maximum Likelihood (cML) and Mantel–Haenszel (M–H) point estimates, derived solely from the time-matched comparisons. The large multiplicative margins of errors, relative to those in Figure 1, are a result of the small size of the sampled risk sets, several of which are concordant (non-informative). The parametric and non-parametric fits to the *aggregated* base series sample provide exposure prevalence estimates that are much more stable, and, thus, a much more precise estimate of the rate ratio. The adjusted MME is only slightly larger than the one based on the variance in Section 2.1, where exposure prevalence was taken to be known.

2.5. *Considerations Due to Further Matching Factors*

When further potential confounders x need to be taken into account, we are still breaking the individual level matching, but may now poststratify with respect to the x-covariates, creating the strata $k = 1, \dots, p$. Letting $s(x_i) \in \{1, \dots, p\}$ indicate the covariate stratum of individual i, the conditional likelihood is now of the form

$$L(\theta; \pi) \overset{\theta}{\propto} \prod_{i=1}^{n} \frac{[\theta \hat{\pi}_{s(x_i)}(t_i)]^{Z_i(t_i)}}{1 - \hat{\pi}_{s(x_i)}(t_i) + \theta \hat{\pi}_{s(x_i)}(t_i)}, \qquad (4)$$

where $\hat{\pi}_k(t)$ is the non-parametric estimator of the exposure prevalence function in stratum k. The corresponding M–H estimator is obtained accordingly from (2). However, it is apparent from the form of the resulting estimator that making the stratification very fine eventually cancels the efficiency gain from pooling of the base series data due to the need to estimate k separate vaccination time distributions $\hat{F}_k$ and eventual vaccination prevalences α_k. This illustrates the bias-variance tradeoff inherent to statistical modeling; establishing what this tradeoff is in the present setting is a topic for further research. We note that a variance estimator corresponding to the estimator maximizing expression (4) can be obtained as

$$\begin{aligned}\hat{V}[\hat{\eta}] &= -I^{\eta\eta}(\hat{\eta}; \hat{F}_1, \dots \hat{F}_p, \hat{\alpha}_1, \dots, \hat{\alpha}_p)^{-1} \\ &\quad + \sum_{k=1}^{p} I^{\eta\eta}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)^{-1} I^{\eta F_k}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k) V[\hat{F}_k] \\ &\quad \times I^{\eta F_k}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)' I^{\eta\eta}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)^{-1} \\ &\quad + \sum_{k=1}^{p} I^{\eta\eta}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)^{-1} I^{\eta\alpha_k}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k) V[\hat{\alpha}_k] \\ &\quad \times I^{\eta\alpha_k}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)' I^{\eta\eta}(\hat{\eta}; \hat{F}_k, \hat{\alpha}_k)^{-1}.\end{aligned}$$

Compared to poststratification, a more sensible approach might be to use a semi-parametric modeling approach for the vaccination times themselves; parsimonious parametrizations can then be applied to obtain model-based estimates for the exposure prevalences $\pi(t, x_i)$. For instance, we might take $\pi(t, x_i) = P(A_i = 1 \mid x_i; \alpha, \beta)[F(t \mid x_i, \phi, S_0) - F(t - 7 \mid x_i, \phi, S_0)]$, where $\text{logit}\{P(A_i = 1 \mid x_i; \alpha, \beta)\} = \alpha + \beta' x_i$ and $F(t \mid x_i, \phi, S_0) = 1 - S_0(t)^{\exp(\phi' x_i)}$, and S_0 is a non-parametrically specified baseline survival function for vaccination time, in practice estimated through the Breslow estimator for cumulative baseline hazard (e.g., Kalbfleisch and Prentice, 2002, p. 117). In practice both of these models would be estimated by fitting them to the pooled base series vaccination time and covariate data, with the resulting exposure prevalences $\pi(t_i, x_i; \hat{\alpha}, \hat{\beta}, \hat{\phi}, \hat{S}_0)$ plugged in to (1). We note that when $\beta \to 0$ and $\phi \to 0$, the model reduces to the semi-parametric special case of Section 2.4, though using the Breslow/Nelson–Aalen estimator, rather than Kaplan–Meier, for the vaccination time distribution.

3. Self-Matched Case-Base Sampling

Controlling for confounding as described in Section 2.5 would require modeling of the exposure prevalence conditional on relevant confounders. However, some of these may be unmeasured, or difficult to model, such as neighborhood (which was used as a matching factor by Patel et al., 2011). In contrast, self-matching automatically controls for time-invariant individual level confounders. In a vaccination context it might again be reasonable to assume that the outcome events are generated by a Poisson process, so that the first event does not terminate the follow-up, nor alter the subsequent event rate, and that the events also do not modify the subsequent vaccination rate (Whitaker, Hocine, and Farrington, 2009, p. 11–12). The self-controlled case series method of Farrington (1995) is based on modeling the age effect on outcome incidence using only the cases. Here, we show how self-matching can be carried out through case-base sampling of person-moments as in Hanley and Miettinen (2009), but drawing the base series person-moments only from the population-time contributed by the individuals with an outcome event.

Again, we select the case series to comprise the person-moments corresponding to outcome events at times t_{i1}, $i = 1, \dots, n$. Supposing that each of the n individuals with an outcome event is followed up through the interval $(0, c_i]$, where c_i may not depend on the outcome, the base series is ascertained by randomly sampling m_i person-moments at times $t_{i2}, \dots, t_{i(m_i+1)}$, $t_{ij} \in (0, c_i]$ for each $i = 1, \dots, n$. (In our running example, $c_i = \tau$ is constant.) These times are generated by a "sampler" counting process $R_i(t) \sim \text{Poisson}(\Lambda^*(t))$, so that $m_i \sim \text{Poisson}(\Lambda^*(c_i))$. $\Lambda^*(t) \equiv \int_0^t \lambda^*(t)\,\mathrm{d}t$ is a user-specified cumulative hazard function chosen to obtain the desired $E[m_i] = \Lambda^*(c_i)$ and age distribution of the sampled person-moments. In the special case of Hanley and Miettinen (2009), $\lambda^*(t) = \lambda^* = M / \sum_{i=1}^{n} c_i$, where M is the total expected base series size (e.g., $M = 100n$), in which case $m_i \sim \text{Poisson}(M c_i / \sum_{i=1}^{n} c_i)$ and $t_{ij} \sim U(0, c_i)$, $j = 2, \dots, m_i + 1$. This sampling mechanism is illustrated in the schematic of Figure a in the Supplementary Appendix B.

In the Supplementary Appendix B we show that, with the assumptions usually made in the self-matched context, by sampling of the base series with uniform probabilities and assuming the proportional hazards model $P(\mathrm{d}N_i(t) = 1 \mid N_i(t-), Z_i(t))/\mathrm{d}t = \lambda_{Z_i(t)}(t, \alpha_i) = \exp\{\alpha_i + f(t, \beta) + \eta Z_i(t)\}$, the conditional likelihood contribution for the $m_i + 1$ person-moments contributed by an individual i with a single outcome event is

$$\frac{\exp\{f(t_{i1}, \beta) + \eta Z_i(t_{i1})\}}{\sum_{j=1}^{m_i+1} \exp\{f(t_{ij}, \beta) + \eta Z_i(t_{ij})\}}. \qquad (5)$$

The individual level intercept terms α_i canceled out, illustrating the effect of self-matching. This expression is readily interpretable as the probability of the event occurring at

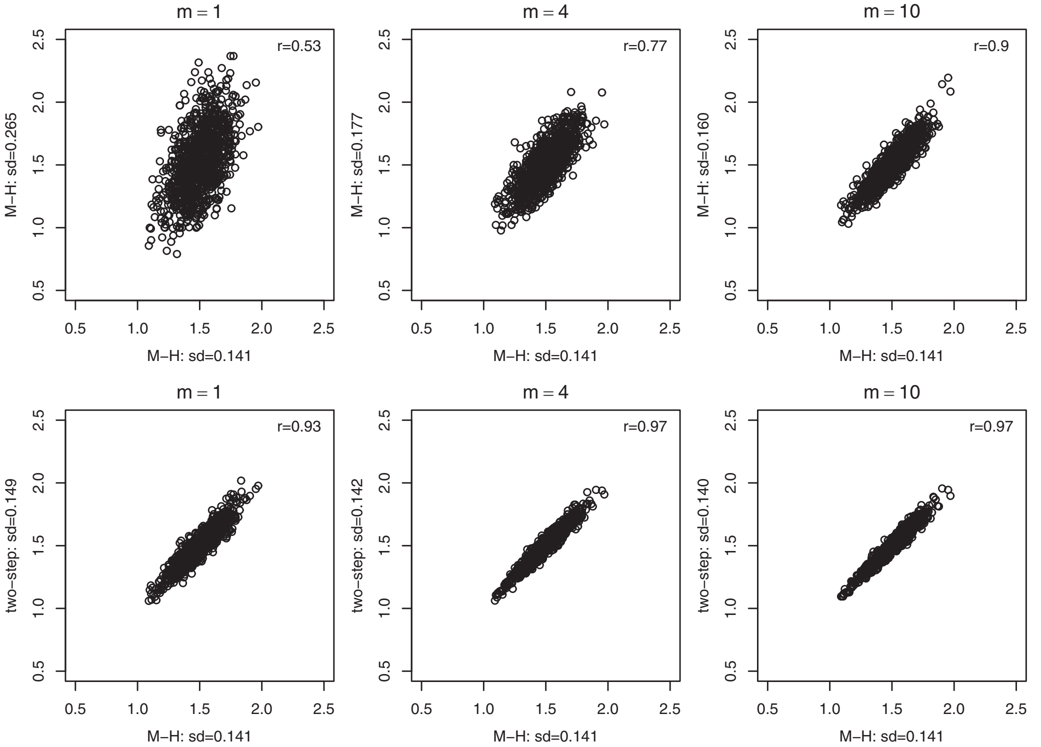


Figure 3. Variation and covariation in 1000 estimates of a rate ratio when the daily exposure prevalences were known (x-axis) or estimated (y-axis) from base-series that were $m = 1$ (left), 4 (middle), and 10 (right) times the size of the case series, using m separate controls for each case (upper panels) or pooling of the controls (lower panels). The true rate ratio was set at exp{1.5}, and the expected (null) number of cases was set at 300. The timing and uptake of vaccination were the same as in Figure 1.

time t_{i1}, given that we know that one of the $m_i + 1$ sampled person-moments involves an event. Also, since (5) is a conditional likelihood in the usual sense (of Cox and Hinkley, 1974, pp. 16–17), variance estimates may be obtained by inverting the observed information matrix at the maximum likelihood point. It should be noted that the person-moment sampling approach of Hanley and Miettinen (2009) operates in continuous time without the need to discretize or split the time axis, and thus (5) can be seen as a continuous, smooth version of the piecewise constant formulation of Farrington (1995,p. 230). However, simple parametric functions of age, such as the linear ($f(t, \beta) \equiv \beta t$) or quadratic ($f(t, \beta) \equiv \beta_1 t + \beta_2 t^2$) functions, are unlikely to adequately control for confounding due to age. Because of this, Ghebremichael-Weldeselassie, Whitaker, and Farrington (2014) suggested using a smooth flexible function to capture the age effect. We also propose using a spline to estimate the function f, but note that, unlike the continuous-time likelihood expression of Ghebremichael-Weldeselassie et al. (2014), the sampling-based likelihood expression (5) does not feature an integral in the denominator, and thus can be easily fitted using standard conditional logistic regression software and any appropriate regression spline.

4. Simulation Study on the Efficiency Gains

4.1. *Time-Matched and Two-Step Estimation*

To demonstrate the efficiency gains of the proposed approach compared to standard time-matched analysis, we simulated outcome and exposure realizations from the data generating mechanism illustrated in Figure 1, using either 1, 4, or 10 controls per case, or 50 controls overall independently of the number of cases. In the latter setting we are interested in how the proposed adjusted variance estimators for the parametric and semi-parametric two-step estimators perform when the overall number of controls is very small. The events were generated from a non-homogeneous Poisson process, but to check the sensitivity to this assumption, only the first event of each individual was used in fitting of the models. For the parametric estimator we fit the gamma distribution for vaccination time to serve as a benchmark for the semi-parametric estimator that uses the ECDF. The null number of cases was set to 300 in a population of 10,000 individuals followed up for 20 weeks from birth, with rate ratios $\theta = 0$, exp{0.5} and exp{1.5}. The potential maximal efficiency gain to be obtained from pooling of the controls can be illustrated by considering the correlation of the M–H estimator (2), in which the true exposure prevalences are assumed to be known, to the counterpart with the unknown $\pi(t, x_i)$s replaced by $H_1(t)/m$. This comparison is presented in the top row of Figure 3; when only one control is selected per case, the correlation is only 0.53, while selecting 10 controls per case increases this to 0.90. This can be contrasted to comparing the M–H estimator with true exposure prevalences to the semi-parametric two-step estimator that uses data from all controls (bottom row). Now selecting only one control per case and using the vaccination data from all controls realizes already almost all of the potential

Table 1

Results for point estimators of $\eta = \log(\theta)$ and estimated standard errors of $\hat{\eta} = \log(\hat{\theta})$. The numbers are means (standard deviations) over 1000 replications. M is the total base series size.

		Estimator					
		π Known		π Unknown			
η	M	M–H (2)	cML (1)	M–H	SE (RBG)	Cond. logistic	SE
0.0	$10n$	−0.021 (0.242)	−0.021 (0.242)	−0.021 (0.253)	0.251 (0.022)	−0.022 (0.252)	0.251 (0.022)
	$4n$	−0.021 (0.242)	−0.021 (0.242)	−0.016 (0.268)	0.268 (0.023)	−0.016 (0.266)	0.268 (0.022)
	n	−0.021 (0.242)	−0.021 (0.242)	−0.013 (0.352)	0.342 (0.032)	−0.013 (0.352)	0.342 (0.032)
0.5	$10n$	0.492 (0.194)	0.492 (0.194)	0.495 (0.207)	0.209 (0.013)	0.494 (0.206)	0.208 (0.013)
	$4n$	0.492 (0.194)	0.492 (0.194)	0.501 (0.223)	0.227 (0.014)	0.500 (0.222)	0.226 (0.014)
	n	0.492 (0.194)	0.492 (0.194)	0.516 (0.304)	0.306 (0.028)	0.516 (0.304)	0.306 (0.028)
1.5	$10n$	1.496 (0.141)	1.495 (0.139)	1.498 (0.160)	0.159 (0.007)	1.496 (0.157)	0.155 (0.006)
	$4n$	1.496 (0.141)	1.495 (0.139)	1.502 (0.177)	0.181 (0.009)	1.501 (0.173)	0.176 (0.008)
	n	1.496 (0.141)	1.495 (0.139)	1.523 (0.265)	0.267 (0.028)	1.523 (0.265)	0.267 (0.028)

		Estimator					
η	M	Parametric Two-step	Naive SE	Adjusted SE	Semi-par. two-step	Naive SE	Adjusted SE
0.0	$10n$	−0.021 (0.242)	0.239 (0.022)	0.240 (0.022)	−0.022 (0.242)	0.239 (0.022)	0.240 (0.022)
	$4n$	−0.021 (0.242)	0.239 (0.022)	0.241 (0.022)	−0.021 (0.242)	0.239 (0.022)	0.241 (0.022)
	n	−0.018 (0.246)	0.239 (0.022)	0.244 (0.022)	−0.019 (0.246)	0.240 (0.022)	0.245 (0.022)
	50	−0.013 (0.274)	0.240 (0.023)	0.268 (0.022)	−0.026 (0.272)	0.243 (0.023)	0.272 (0.022)
0.5	$10n$	0.492 (0.194)	0.195 (0.013)	0.196 (0.013)	0.492 (0.194)	0.195 (0.013)	0.196 (0.013)
	$4n$	0.492 (0.195)	0.195 (0.013)	0.197 (0.013)	0.492 (0.195)	0.195 (0.013)	0.197 (0.013)
	n	0.495 (0.197)	0.195 (0.013)	0.201 (0.013)	0.495 (0.198)	0.196 (0.013)	0.202 (0.013)
	50	0.506 (0.229)	0.197 (0.013)	0.231 (0.015)	0.497 (0.233)	0.199 (0.014)	0.235 (0.015)
1.5	$10n$	1.494 (0.140)	0.140 (0.005)	0.141 (0.005)	1.495 (0.140)	0.140 (0.005)	0.141 (0.005)
	$4n$	1.495 (0.141)	0.140 (0.005)	0.142 (0.005)	1.495 (0.142)	0.140 (0.005)	0.143 (0.005)
	n	1.498 (0.148)	0.141 (0.005)	0.148 (0.006)	1.498 (0.149)	0.141 (0.005)	0.149 (0.006)
	50	1.516 (0.190)	0.142 (0.008)	0.194 (0.018)	1.529 (0.197)	0.144 (0.008)	0.198 (0.018)

improvement, with a correlation of 0.93; in other words, the two-step estimator with *one control* per case gives improved efficiency over standard matched analysis with *10 controls* per case.

Numerical results for various point estimators and selected variance estimators are presented in Table 1. The estimated standard errors should be compared with the Monte Carlo (MC) standard deviations of the corresponding point estimators. The M–H and conditional logistic estimator with the exposure prevalences unknown correspond to standard matched analyses of case-control data. The MC standard deviations in Table 1 again indicate that with one or four controls per case, the semi-parametric estimator gives a clear efficiency gain compared to matched analysis, and comes close in efficiency to knowing the true exposure prevalences. It is also notable that using the parametric two-step approach by fitting the correctly specified vaccination time density gives only very minor improvements compared to the semi-parametric approach. For two-step estimation, Table 1 presents both the naive variance estimators, which ignore the estimation of nuisance parameters, and the adjusted versions. With one or four controls selected per case the adjustment has a rather negligible effect on the estimated standard errors, indicating that in these scenarios adjusting the variance is hardly necessary. However, when only 50 controls are selected overall, the naive standard errors are clearly too low, while the adjusted ones match to the MC standard deviations.

4.2. *Self-Matched Estimation*

We fitted the self-matched conditional likelihood (5) by sampling uniformly (in expectation) 10, 40, or 100 base series person-moments per case from the follow-up time contributed by the individuals with an outcome event. A quadratic function of age, as well as a cubic spline basis (the default option in R `bs` function), were fitted to account for confounding by age. The results are shown in Table 2. The self-matched case-base sampling resulted in very good efficiency even with only 40 base series person-moments per case, but the quadratic function of age was clearly not sufficient for controlling for the age effect. This demonstrates the need for more flexible semi-parametric modeling in controlling for the age effect; in contrast, the low-dimensional regression spline does much better in terms of bias, without inflating the standard errors too much.

Table 2

Results for the self-matched point-estimators of $\eta = \log(\theta)$ and the corresponding standard errors. M is the expected total number of self-matched base series person-moments.

		Estimator			
η	M	Self-matched (quadratic)	SE	Self-matched (spline)	SE
0.0	$100n$	0.097 (0.242)	0.238 (0.022)	−0.012 (0.245)	0.242 (0.022)
	$40n$	0.098 (0.246)	0.242 (0.022)	−0.012 (0.250)	0.246 (0.022)
	$10n$	0.113 (0.271)	0.264 (0.021)	−0.005 (0.275)	0.268 (0.021)
0.5	$100n$	0.612 (0.191)	0.194 (0.013)	0.504 (0.194)	0.199 (0.012)
	$40n$	0.612 (0.198)	0.199 (0.012)	0.503 (0.201)	0.203 (0.012)
	$10n$	0.617 (0.222)	0.223 (0.012)	0.498 (0.225)	0.228 (0.012)
1.5	$100n$	1.621 (0.140)	0.141 (0.005)	1.513 (0.146)	0.147 (0.005)
	$40n$	1.620 (0.146)	0.146 (0.005)	1.511 (0.154)	0.152 (0.005)
	$10n$	1.639 (0.170)	0.174 (0.007)	1.521 (0.176)	0.180 (0.007)

5. Discussion

Compared to standard time-matched analysis through the Mantel–Haenszel method or conditional logistic regression, where the riskset at each event time involves only the sampled controls as well as the case itself, pooling of the controls does result in improved efficiency; this is evident also in the setting of the present article (Section 4). The efficiency gain is most strikingly illustrated in Figure 3, where the Mantel–Haenszel estimates from matched case-control sets are correlated with those using the time-specific exposure prevalence estimated from the pooled control data. However, compared to the pooling approach discussed by Samuelsen (1997), instead of using sampling weights, our outlined approach is based on estimating the population exposure prevalence, which does not require an enumerable study population, or determination of the inclusion probabilities in the sampling scheme.

The self-matched case-base sampling approach using a homogeneous Poisson process to sample the base series person-moments resulted in a likelihood expression that can be easily fitted using standard conditional logistic regression. Drawing the self-matched base series with informative probabilities is a topic for further work; we aim to investigate two-step estimation approaches that would use the non-parametrically estimated exposure prevalence function (Section 2.4) in the specification of the function $\lambda^*(t)$ (Section 3). Using this additional information in sampling of the self-matched base series through a non-homogeneous Poisson process could potentially improve the efficiency of the sampling and also remove most or all of the multiplicative age effect in the process. However, in the simulation study of Section 4 we only appled the self-matching with a uniformly sampled base series to establish a proof of concept.

We proposed also a way to illustrate the pooled exposure data, available on either the whole study population or a control/base series, in a two-dimensional (population-time) plot (Figures 1 and 2). A conventional graphical display for temporal association of the vaccination time/exposure and incident cases is obtained when the two-dimensional presentation is collapsed over the population dimension into the time dimension. The population-time plot allows for direct visual comparison of incidence density in the exposed versus unexposed population time, which is lost in the temporal association presentation.

6. Supplementary Materials

Supplementary Web Appendices, referenced in Sections 2.1 2.3, 2.4 and 3, as well as the R code for producing Figures 1–3 and the simulation results, are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The majority of this work was carried out when the first author was working at the Department of Epidemiology, Biostatistics and Occupational Health of McGill University. This work was supported by the Finnish Foundation for Technology Promotion (to O.S.) and the Natural Sciences and Engineering Research Council of Canada (NSERC) (to O.S. and J.H.).

References

Aalen, O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View.* New York: Springer.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34**, 86–102.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **51**, 228–235.

Farrington, C. P. and Whitaker, H. J. (2006). Semiparametric analysis of case series data. *Journal of the Royal Statistical Society, Series C* **55**, 553–594.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Self-controlled case series method with smooth age effect. *Statistics in Medicine* **33**, 639–649.

Gray, R. J. (2009). Weighted analyses for cohort sampling designs. *Lifetime Data Analysis* **15**, 24–40.

Greenland, S. (1999). A unified approach to the analysis of case-distribution (case-only) studies. *Statistics in Medicine* **18**, 1–15.

Hanley, J. A. and Miettinen, O. S. (2009). Fitting smooth-in-time prognostic risk functions via logistic regression. *The International Journal of Biostatistics* **5**(1), Article 3.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. New Jersey: Wiley.

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiological cohort studies. *Statistical Science* **11**, 35–53.

Månsson, R., Joffe, M. M., Sun, W., and Hennessy, S. (2007). On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology* **166**, 332–339.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

Miettinen, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology* **103**, 226–235.

Miettinen, O. S. and Karp, I. (2012). *Epidemiological Research: An Introduction*. Dordrecht: Springer.

Nohynek, H., Jokinen, J., Partinen, M., Vaarala, O., Kirjavainen, T., Sundman, J., Himanen, S. L., Hublin, C., Julkunen, I., Olsen, P., Saarenpaa-Heikkila, O., and Kilpi, T. (2012). AS03 adjuvanted AH1N1 vaccine associated with an abrupt increase in the incidence of childhood narcolepsy in Finland. *PLoS ONE* **7**, e33536.

Patel, M. M., López-Collada, V. R., Bulhões, M. M., De Oliveira, L. H., Bautista Márquez, A., Flannery, B., Esparza-Aguilar, M., Montenegro Renoiner, E. I., Luna-Cruz, M. E., Sato, H. K., Hernández-Hernández, L. del C., Toledo-Cortina, G., Cerón-Rodríguez, M., Osnaya-Romero, N., Martínez-Alcazar, M., Aguinaga-Villasenor, R. G., Plascencia-Hernández, A., Fojaco-González, F., Hernández-Peredo Rezk, G., Gutierrez-Ramírez, S. F., Dorame-Castillo, R., Tinajero-Pizano, R., Mercado-Villegas, B., Barbosa, M. R., Maluf, E. M., Ferreira, L. B., de Carvalho, F. M., dos Santos, A. R., Cesar, E. D., de Oliveira, M. E., Silva, C. L., de Los Angeles Cortes, M., Ruiz Matus, C., Tate, J., Gargiullo, P., and Parashar, U. D. (2011). Intussusception risk and health benefits of rotavirus vaccination in Mexico and Brazil. *New England Journal of Medicine* **364**, 2283–2292.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**, 311–323.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **6**, 41–55.

Saarela, O., Kulathinal, S., Arjas, E., and Läärä, E. (2008). Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in Medicine* **27**, 5991–6008.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**, 379–394.

Samuelsen, S. O., Ånestad, H., and Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics* **34**, 103–119.

Silcocks, P. (2005). An easy approach to the Robins-Breslow-Greenland variance estimator. *Epidemiologic Perspectives & Innovations* **2**, 9.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56**, 29–38.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Whitaker, H. J., Hocine, M. N., and Farrington, C. P. (2009). The methodology of self-controlled case series studies. *Statistical Methods in Medical Research* **18**, 7–26.

Received August 2013. Revised July 2014. Accepted July 2014.