HHS Public Access

Author manuscript

Biom J. Author manuscript; available in PMC 2019 June 24.

Published in final edited form as:

Biom J. 2018 July ; 60(4): 748–760. doi:10.1002/bimj.201700088.

Analyzing self-controlled case series data when case confirmation rates are estimated from an internal validation sample

Stanley Xu[1,2], **Christina L. Clarke**[1], **Sophia R. Newcomer**[1,2], **Matthew F. Daley**[1,3], and **Jason M. Glanz**[1,2]

[1]The Institute for Health Research, Kaiser Permanente Colorado, Denver, CO 80231, USA

[2]School of Public Health, University of Colorado, Aurora, CO 80045, USA

[3]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA

Abstract

Vaccine safety studies are often electronic health record (EHR)-based observational studies. These studies often face significant methodological challenges, including confounding and misclassification of adverse event. Vaccine safety researchers use self-controlled case series (SCCS) study design to handle confounding effect and employ medical chart review to ascertain cases that are identified using EHR data. However, for common adverse events, limited resources often make it impossible to adjudicate all adverse events observed in electronic data. In this paper, we considered four approaches for analyzing SCCS data with confirmation rates estimated from an internal validation sample: (1) *observed cases*, (2) *confirmed cases* only, (3) *known confirmation rate*, and (4) *multiple imputation* (MI). We conducted a simulation study to evaluate these four approaches using type I error rates, percent bias, and empirical power. Our simulation results suggest that when misclassification of adverse events is present, approaches such as *observed cases, confirmed case only*, and *known confirmation rate* may inflate the type I error, yield biased point estimates, and affect statistical power. The *multiple imputation* approach considers the uncertainty of estimated confirmation rates from an internal validation sample, yields a proper type I error rate, largely unbiased point estimate, proper variance estimate, and statistical power.

Keywords

confirmation rate of cases; internal validation sample; multiple imputation; self-controlled case series; vaccine safety

Correspondence Stanley Xu, The Institute for Health Research, Kaiser Permanente Colorado, 10065 E. Harvard Avenue, Denver, CO 80231, USA. stan.xu@kp.org.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

1 | BACKGROUND

Observational studies using electronic health record (EHR) data are important in health care research and in postmarketing surveillance of drug and vaccine safety (Hersh, 2007; Platt & Carnahan, 2012; Platt et al., 2009; Safran et al., 2007; Weiskopf & Weng, 2013). Using EHR data, vaccine safety studies from the Vaccine Safety Datalink (VSD) and Postlicensure Rapid Immunization Safety Monitoring (PRISM) have helped inform national immunization policies (Baggs et al., 2011; Curtis et al., 2012; McNeil et al., 2014; Nguyen, Ball, Midthun, & Lieu, 2012). Like other observational studies, VSD and PRISM studies face significant methodological challenges since EHR data are collected as part of routine clinical care, and are then repurposed to conduct studies of adverse events related to vaccinations. Confounding and misclassification of adverse event are two methodological challenges when analyzing EHR data for vaccine safety research.

The self-controlled case series (SCCS) method was developed for vaccine safety studies to handle confounding (Farrington, 1995). The SCCS method is limited to only those individuals who experience adverse events (i.e., cases) in an exposed interval (i.e., risk interval) and/or unexposed intervals (i.e., control interval). By comparing the incidence rate of medically attended events in the exposed interval following a vaccination to the incidence rate in the unexposed intervals within an individual, SCCS design adjusts for the effects of time-invariant confounders by allowing each person to act as their own control. Primarily due to this advantage, the SCCS design has been widely used in vaccine safety studies (Hambidge et al., 2006; Miller, Waight, Farrington, Stowe, & Taylor, 2001; Stowe, Andrews, Ladhani, & Miller, 2016; Sun et al., 2012).

In typical vaccine safety studies using a SCCS design, presumptive adverse events are first identified by the presence of medical codes, such as *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9) codes or recent *tenth revision* ICD-10 codes in EHR data. Case misclassification occurs when the codes obtained from EHR data do not represent true cases of disease, such as when the diagnosis is miscoded, or when the diagnosis code was applied to a patient's medical record but the medical condition was later ruled-out (Chubak, Pocobelli, & Weiss, 2012). To reduce misclassification bias, researchers will review the clinical provider notes of presumptive adverse events identified in EHR data. After this medical chart review, a positive confirmation rate (PCR) can be calculated as the proportion of adverse events confirmed among those presumptive adverse events reviewed with medical charts (Xu et al., 2014). As a measure of medical coding accuracy, the PCR will vary based on the type of adverse event, setting (e.g., inpatient hospitalization, emergency visit, primary care visit, etc.), and will also vary across different managed care organizations (MCOs) (Mullooly et al., 2004; Mullooly, Donahue, DeStefano, Baggs, & Eriksen, 2008). The PCR previously observed in vaccine safety studies has ranged between 6.8% and 95.0% in VSD data (McNeil et al., 2016; Mullooly et al., 2008).

Many previous vaccine safety studies have used the SCCS design to study severe and rare adverse events. In these studies, researchers typically conduct chart reviews of all presumptive cases identified in risk and control intervals, and then reanalyze the data with only confirmed cases. However, more common adverse events have also been studied (Glanz

et al., 2016; Hambidge et al., 2014). When the number of adverse events identified by medical coding is relatively large and resources to review all the presumptive cases are constrained, researchers may only be able to review a sample of identified adverse events (internal validation sample). For these situations, methods are needed to inform how to apply information from a PCR based on just a sample of presumptive cases in examining the association between vaccination and adverse events.

Recently, Xu et al. (2014) developed a statistical model for analyzing SCCS data with an imperfect PCR that was considered as a known parameter without uncertainty. However, the PCR estimated from an internal validation sample has variance and researchers should consider this uncertainty in applying the PCR to those cases not reviewed. In addition, outcome (case) misclassification has sometimes been treated as a missing data problem (Cole, Chu, & Greenland, 2006; Edwards, Cole, Troester, & Richardson, 2013; Lyles et al., 2011), and statistical methods have been developed by applying techniques for missing data such as multiple imputation (*MI*) (Rubin, 1987). In vaccine safety studies using EHR data, the lack of case confirmation information among those cases not reviewed can also be considered as a missing data problem. If cases for medical chart review are randomly sampled, then the absence of a case confirmation can be considered as missing at random (MAR), meeting the assumption of the *MI* approach.

In this paper, we considered four approaches for analyzing SCCS data with imperfect confirmation rates estimated from an internal validation sample: (1) an *observed cases* approach that analyzes presumptive cases without medical chart review, (2) a *confirmed cases only* approach that analyzes only confirmed cases from a sample of observed cases, (3) a *known confirmation rate* approach that estimates the PCR from a sample of adverse events, excludes those that are reviewed and not confirmed, and applies the estimated PCR to those that are not reviewed without considering the uncertainty of PCR estimates, and (4) a *multiple imputation* approach that estimates the PCR, excludes those that are not confirmed, and applies the estimated PCR to those that are not reviewed using a *MI* approach to take the uncertainty of PCR estimates into account. We conducted a simulation study to evaluate these four approaches using type I error rates, percent bias, Monte Carlo error (MCE), and empirical power.

2 | STATISTICAL METHODS

2.1 | Confirmation rate of electronically identified adverse events when only some of adverse events were reviewed with medical chart

When planning chart reviews of relatively common adverse events, researchers often consider the number of cases and exposure status of cases. In a SCCS design, an individual's follow-up time is partitioned into a risk interval that occurs immediately after a vaccine exposure, and control intervals occurring prior to vaccination and/or after the end of the risk interval (Farrington, 1995; Xu et al., 2011; Xu, Hambidge, McClure, Daley, & Glanz, 2013). Cases that occur during the risk interval are called exposed cases and those occurring during a control interval are called unexposed cases. In multisite research, such as with the VSD, the participating sites where the cases are identified are also considered. If a certain number of cases (internal validation sample) are selected completely at random, study sites with

larger patient population will incur a larger burden of chart reviews while smaller study sites may not have any sampled cases to review. In addition, since the risk interval is usually shorter than the control interval, the number of unexposed cases is generally much larger than the number of exposed cases. Thus, researchers tend to review fewer charts from unexposed cases.

Let k denote the study site and e denote the exposure status with $e = 1$ for exposed cases and $e = 0$ for unexposed cases. Also, let $\boldsymbol{T}_{k1}$ and $\boldsymbol{T}_{k0}$ denote the total number of presumptive adverse events based on medical coding in the risk and control intervals for site k. Assume $\boldsymbol{G}_{ke}$ out of $\boldsymbol{T}_{ke}$ cases are reviewed and g_{ke} cases are confirmed after medical chart review. Let R_{k1} and R_{k0} denote the proportions of reviewed cases and let R_{k1} and R_{k0} denote the confirmation rates for exposed and unexposed cases for site k, respectively. The proportion of reviewed cases is $\mathbf{R}_{ke} = \frac{G_{ke}}{T_{ke}}$, and the confirmation rate can be estimated as follows:

$$\tilde{q}_{ke} = \frac{g_{ke}}{\mathbf{G}_{ke}} \quad (1)$$

In our simulation study, we use k = S, M, and L to represent small, medium, and large study sites.

2.2 | Conditional Poisson models for analyzing SCCS data with imperfect PCR estimated from an internal validation sample

We considered four methods to analyze SCCS data, including a method accounting for the uncertainty of the estimated internal validation sample's PCR. As in Xu et al. (2014), let y_{ij} denote the observed number of presumptive adverse events based on medical coding in period j for individual i. Among the total number of observed presumptive adverse events, let y_{ijc} present the number of correctly classified adverse events and let y_{ijw} represent the number of misclassified adverse events, so that $y_{ij} = y_{ijc} + y_{ijw}$. Let p_{ij} represent the PCR of a presumptive case based on the accuracy of medical coding in EHR data. The value of p_{ij} for a case can be assigned by $\tilde{q}_{ke}$ depending on the study site and exposure status of the case. For example, if a case is from site $k = L$ and period j has an exposure status $e = 1$, then $p_{ij} = \tilde{q}_{L1}$. We evaluated the following four approaches for estimating the vaccination effect when confirmation rates of observed cases are estimated from an interval validation sample:

2.2.1 | Observed cases approach that ignores misclassification—

Assuming there is no misclassification of cases, Farrington proposed a conditional Poisson model (1995) to analyze SCCS data conditional on the total number of observed adverse events:

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left[\frac{t_{ij} \exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij} \exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}} \quad (2)$$

where N is the number of individuals who experienced adverse events and n_i is the number of time intervals for individual i. Also t_{ij} is person time (e.g., days), $\boldsymbol{X}_{ij}$ is the row vector of time-varying covariates including the vaccination status (x_1) and other indicator variables for age groups and seasonality. The column vector $\boldsymbol{\beta}$ contains corresponding coefficients β_1, β_2, β_3,…, where β_1 is the coefficient for the vaccination effect and the incidence rate ratio (IRR) is defined by IRR = exp(β_1), and β_2, β_3, … are the coefficients for other covariates such as age groups and seasonality. Maximum likelihood estimates (MLEs) can be obtained by maximizing the log conditional likelihood function as shown in Equation (3) for the conditional Poisson model (2),

$$\ln L_1(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(y_{ij}\right) \left[\ln\left(t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)\right) - \ln \sum_{j=1}^{n_i} \left(t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)\right) \right] \quad (3)$$

2.2.2 | Confirmed cases only approach—Let R indicate whether a case underwent medical chart review ($R = 1$ for reviewed and $R = 0$ for not reviewed) and $\boldsymbol{C}$ indicate whether a reviewed case was confirmed ($\boldsymbol{C} = 1$ for confirmed case and $\boldsymbol{C} = 0$ for an unconfirmed case). A conditional Poisson model including only those cases that were reviewed and confirmed is:

$$L_2(\boldsymbol{\beta}) = \prod_{\boldsymbol{R}_{ij}=1} \prod_j \left[\frac{t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}C_{ij}} \quad (4)$$

2.2.3 | Known confirmation rate approach—We estimated the confirmation rate and then applied it as a known parameter without uncertainty to those cases that were not reviewed. Cases that were reviewed but were unconfirmed were excluded. A modified likelihood function across individuals is:

$$L_3(\boldsymbol{\beta}) = \prod_{\boldsymbol{R}_{ij}=1} \prod_j \left[\frac{t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}C_{ij}} \prod_{\boldsymbol{R}_{ij}=0} \prod_j \left[\frac{t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}p_{ij}} \quad (5)$$

where $\prod_{\boldsymbol{R}_{ij}=1} \prod_j [\frac{t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}{\sum_j t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}]^{y_{ij}C_{ij}}$ is the likelihood contribution from confirmed cases among those reviewed, and $\prod_{\boldsymbol{R}_{ij}=0} \prod_j [\frac{t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}{\sum_j t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}]^{y_{ij}p_{ij}}$ is a pseudo-likelihood contributed from cases that were not reviewed. Here p_{ij} is equal to $\tilde{q}_{ke}$ where case i is from site k and j period has an exposure status e, and $\tilde{q}_{ke}$ is estimated from those reviewed. Thus, this approach allows the PCR to differ between exposed and unexposed cases. Note that, in the likelihood expression (5), p_{ij} is considered as a known parameter without uncertainty

(Xu et al., 2014). Maximum likelihood estimates (MLE) can be obtained by maximizing likelihood functions $L_1(\boldsymbol{\beta})$, $L_2(\boldsymbol{\beta})$, and $L_3(\boldsymbol{\beta})$. Approximate standard errors for the MLEs were computed using the delta method. We calculated the likelihood functions, obtained MLEs and their standard errors in SAS PROC NLMIXED. MLEs and their standard errors were used in calculating type I error and empirical power.

2.2.4 | *Multiple imputation* (MI) approach—We estimated the confirmation rate and then applied it as a parameter with uncertainty to those not reviewed. Cases that were reviewed but were unconfirmed were excluded. If $\boldsymbol{G}_{ke} < \boldsymbol{T}_{ke}$, the uncertainty of $\tilde{q}_{ke}$ exists and should be considered in the likelihood function calculation. Now we introduce a multiple imputation approach to estimate the vaccination effect in a SCCS design while considering the uncertainty of confirmation rates estimated from an interval validation sample. Similarly, the *MI* approach also allows the PCR to differ between exposed and unexposed cases. To achieve this goal, five steps were taken:

1. A logistic regression model with the interaction between site and exposure status for case confirmation among those being reviewed was fit:

$$\text{Logit}(\text{prob}(C = 1 \mid k, e) = \pi_{ke}$$

   where k indicated site and e indicated exposure status. In simulation study, we created six dummy variables to represent site (3 levels) and exposure status (2 levels) for a case and these dummy variables were included in the logistic regression model; $\boldsymbol{\pi}_{ke}$ is the coefficient for site k and exposure status e to be estimated. Theoretically, other factors that may influence misclassification of cases can be included in the probabilistic model.

2. Coefficients and their variances and covariance were obtained from the probabilistic model in step 1: a vector $\hat{\boldsymbol{\pi}}$ including $\hat{\pi}_{\text{s0}}$, $\hat{\pi}_{\text{s1}}$, $\hat{\pi}_{\text{M0}}$, $\hat{\pi}_{\text{M1}}$, $\hat{\pi}_{\text{L0}}$, and $\hat{\pi}_{\text{L1}}$ for sites $k = \boldsymbol{S, M, L}$ and exposure status $e = 0, 1$; covariance matrix $\hat{\Sigma}$ among $\hat{\pi}_{\text{s0}}$, $\hat{\pi}_{\text{s1}}$, $\hat{\pi}_{\text{M0}}$, $\hat{\pi}_{\text{M1}}$, $\hat{\pi}_{\text{L0}}$, and $\hat{\pi}_{\text{L1}}$ was also obtained.

3. For those cases that were not medically chart reviewed ($R = 0$) in site k with exposure status e, random values of $\boldsymbol{\pi}_{\text{s0}}$, $\boldsymbol{\pi}_{\text{s1}}$, $\boldsymbol{\pi}_{\text{M0}}$, $\boldsymbol{\pi}_{\text{M1}}$, $\boldsymbol{\pi}_{\text{L0}}$, and $\boldsymbol{\pi}_{\text{L1}}$ were drawn from a multivariate normal with a vector of means equal to $\hat{\boldsymbol{\pi}}$ and the covariance matrix equal to $\hat{\Sigma}$. Let $\boldsymbol{\varepsilon}_{ij}$ denote the random value of $\boldsymbol{\pi}_{ke}$ for case i in k site and in interval $j = e$.

4. Then an indicator variable Z_{ij} was created by randomly drawing from a Bernoulli distribution with probability p_{ij}

$$p_{ij}(\text{Z}_{ij} = 1 \mid \text{site, exposure}) = \frac{\exp(\varepsilon_{ij})}{1 + \exp(\varepsilon_{ij})}$$

5. The likelihood function across all cases is

$$L_4(\boldsymbol{\beta}) = \prod_{\boldsymbol{R}_{ij}=1} \prod_j \left[\frac{t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}C_{ij}} \prod_{\boldsymbol{R}_{ij}=0} \prod_j \left[\frac{t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)}{\sum_j t_{ij}\exp\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}\right)} \right]^{y_{ij}Z_{ij}} \quad (6)$$

where $\prod_{\boldsymbol{R}_{ij}=1} \prod_j [\frac{t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}{\sum_j t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}]^{y_{ij}C_{ij}}$ is the likelihood contribution from confirmed cases among those that were reviewed, and $\prod_{\boldsymbol{R}_{ij}=0} \prod_j [\frac{t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}{\sum_j t_{ij}\exp(\boldsymbol{X}_{ij}\boldsymbol{\beta})}]^{y_{ij}Z_{ij}}$ is the likelihood contribution from cases that were not reviewed. Steps 3–5 were repeated 100 times for each SCCS dataset in our simulation. The point estimate of $\hat{\beta}_1$ is the average of $\hat{\beta}_1$s from 100 replications and the standard error of $\hat{\beta}_1$ is the square root of the sum of the within-imputation variance and the between-imputation variance from the 100 replications (Rubin, 1987). Point estimates of $\hat{\beta}_1$ and their standard errors were used in calculating type I error and empirical power.

3 | SIMULATION STUDY

3.1 | Simulation algorithm and analyses

The processes for simulating SCCS data have been described elsewhere (Xu et al., 2014). Briefly, 100,000 hypothetical individuals from three study sites were assumed to have a follow-up period of 365 days, consisting of a 42-days' risk interval starting on vaccination day and control intervals before vaccination and after the risk interval. Among the 100,000 individuals 10% were from a small site (S), 30% from a medium site (M) and 60% from a large site (L). Vaccination times were assumed to follow a uniform distribution with a range of 1–365 days. Observed presumptive adverse events were simulated per model (7).

$$y_{ij} \sim poisson\left(\frac{\lambda_{ijc}}{p_{ij}}\right), \lambda_{ijc} = E\left(y_{ijc}\right) = t_{ij}\exp\left(\beta_0 + x_{1ij}\beta_1\right) \quad (7)$$

We simulated 24 scenarios of SCCS studies with different baseline adverse event rates, confirmation rates, proportions of reviewed cases, and vaccination effects. Two values of β_0, −13 and −12, were used to achieve relatively common and rare baseline adverse events, yielding average baseline incidence rates 6.1 cases and 2.3 cases per one million person days, respectively. The coefficient for the true vaccination effect, β_1, was chosen to be 0.69 to represent an IRR of 2. We also established simulations with $\beta_1 = 0$ to study the type I error rate. Table 1 shows the number of observed cases for different β_0 values (determining the baseline rates of adverse events) and different confirmation rates when β_1= 0.69. The number of cases ranged from two exposed cases from the small site with $\beta_0 = -13$ to 201 unexposed cases from the large site with $\beta_0 = -12$.

We examined both scenarios where the PCR of presumptive cases was the same and when it was different between exposed and unexposed cases. We represented the PCR as q_0 for unexposed cases and q_1 for exposed cases. We chose three pairs of q_0 and q_1: 50% and 50%, 80% and 50%, and 50% and 80%. In these simulations, we assumed that charts of all cases from the small site were reviewed. Let R_{M0}, R_{M1}, R_{L0} and R_{L1} denote the proportions of reviewed cases at the medium and large sites in the unexposed and exposed intervals. The impact of five sets of proportions of reviewed cases was investigated (R_{M0}, R_{M1}, R_{L0}, R_{L1}): (100%,100%,100%,100%), (80%,80%,80%,80%), (50%,50%,50%,50%), (50%,100%,50%, 100%), and (80%,100%,80%,100%). The scenario where all cases were chart reviewed (i.e. (R_{M0}, R_{M1}, R_{L0}, R_{L1}) = (100%,100%,100%,100%)) can be considered the gold standard. Recall that all cases were reviewed for small sites (i.e., (R_{S0}, R_{S1}) = (100%,100%)). Among chart reviewed cases, the confirmed cases were assigned based on the confirmation rates established for each simulation. We assumed that the PCR is the same between reviewed and not reviewed cases in each exposed or unexposed interval for a study site.

For each scenario of β_0 and PCRs and proportions of reviewed cases, we simulated 1000 replications. We then analyzed these simulated datasets using the four analytic approaches described in the Statistical Methods section. The following evaluation metrics were obtained from the 1000 replicas.

3.2 | Evaluation metrics

3.2.1 | Type I error rate—The type I error rate was calculated as the proportion of replicated datasets where a statistical test falsely rejected a true null hypothesis ($\beta_1 = 0$) at a significance level of 0.05.

3.2.2 | Percent bias—We provided the mean point estimate and the mean standard error for the vaccination effect coefficient ($\hat{\beta}_1$) over 1000 replicas. The percent bias was calculated as $100 \times (\text{mean } \hat{\beta}_1 - \text{true } \beta_1)/\text{true } \beta_1$, where true β_1 is the simulated vaccination effect (set to $\beta_1 = 0.69$).

3.2.3 | Monte Carlo error (MCE) of $\hat{\beta}_1$—Monte Carlo error (MCE) of $\hat{\beta}_1$ is the between-simulation variability. MCE approximates the true variation of $\hat{\beta}_1$ (Koehler, Brown, & Haneuse, 2009). The MCEs were calculated as the standard deviation of the point estimate $\hat{\beta}_1$ over 1000 replicas and were compared to the mean standard error of $\hat{\beta}_1$ estimated by each of the four approaches.

3.2.4 | Empirical power—Empirical power was calculated as the proportion of datasets where a statistical test correctly rejected a false null hypothesis when simulations were performed under the true alternative $\beta_1 = 0.69$ (i.e., IRR = 2).

3.3 | Simulation results

The scenario where all presumptive cases were chart reviewed and analyzed with the *confirmed cases only* approach was considered the gold standard. We evaluated the performance of the four approaches by comparing each to this gold standard.

3.3.1 | Type I error rates—Type I error rates are shown in Table 2. When the PCR was the same between the exposed and unexposed cases (i.e., $q_0 = q_1 = 50\%$), the type I error rates from the gold standard were 4.2% and 3.5% for $\beta_0 = -12$ and $\beta_0 = -13$, respectively. The *observed cases* approach analyzing all electronically identified presumptive cases yielded type I error rates close to the nominal value (5%). However, when the PCR was different, the *observed cases* approach yielded type I error rates up to 78.9% with $\beta_0 = -12$, which was far greater than the type I error rates near 5% from the gold standard.

Regardless of whether PCRs were different or the same between exposed and unexposed cases, if the proportions of observed cases reviewed with medical chart were the same for exposed and unexposed cases across sites, the *confirmed cases only* approach yielded type I error rates close to 5%. However, when the proportion of observed cases medically chart reviewed were different by exposure status, the *confirmed cases only* approach inflated the type I error rates up to 71.9%.

The *known confirmation rate* approach yielded a type I error rate close to 5% except when the proportions of reviewed cases were low. When (R_{M0}, R_{M1}, R_{L0}, R_{L1}) = (50%,50%,50%, 50%), $q_0 = 50\%$ and $q_1 = 50\%$, the type I error rates were 11.5% and 9.6% for $\beta_0 = -12$ and $\beta_0 = -13$, respectively. The type I error rates also increased to near 10% when q_0 and q_1 differed. Unlike the other three approaches, the *MI* approach always produced type I error rates between 3.1% and 5.3%.

3.3.2 | Mean (standard error) of $\hat{\beta}_1$, percent bias, and empirical power—The mean and standard error of $\hat{\beta}_1$ and percent bias are presented in Tables 3 and 4 for $\beta_0 = -12$ and $\beta_0 = -13$, respectively. Empirical power is presented in Table 5. When $\beta_0 = -12$ and $q_0 = q_1 = 50\%$, the gold standard yielded a percent bias of –1% with mean $\hat{\beta}_1 = 0.68$ and standard error = 0.17 (Table 3), and an empirical power of 95.9% (Table 5). Compared to the gold standard, the *observed cases* approach yielded smaller standard error of $\hat{\beta}_1$ (0.12 versus 0.17 from the gold standard) (Table 3) and produced greater empirical power (100% versus 95.9% from the gold standard) (Table 5). When $\beta_0 = -12$ and $q_0 \neq q_1$, the *observed cases* approach yielded biased estimates and empirical power deviated from the empirical power established with the gold standard method.

Regardless of whether the PCR was the same or different for exposed and unexposed cases, if the proportion of observed cases medically chart reviewed were the same for exposed and unexposed cases across sites, the *confirmed cases only* approach yielded an unbiased estimate for $\hat{\beta}_1$ but produced a larger standard error as the proportion of medically chart reviewed cases decreased. However, when the proportion of medically chart reviewed cases were different by exposure status, the *confirmed case only* approach overestimated $\hat{\beta}_1$ by up to 87% for $\beta_0 = -12$ and it resulted in inflated empirical power: 100% versus 96.3 % from the *MI* approach for $\beta_0 = -12$.

The *known confirmation rate* approach yielded an unbiased estimate for $\hat{\beta}_1$ regardless of the confirmation rate and the proportion of reviewed cases when cases were common (i.e., $\beta_0 =$

−12). However, when the proportion of reviewed cases were low (e.g., 50%), this method resulted in smaller standard errors and inflated empirical power compared to the *MI* approach, especially when cases were rare (i.e., $\beta_0 = -13$). For example, when $R_{M0} = R_{M1} = R_{L0} = R_{L1} = 50\%$ and $\beta_0 = -13$, the *known confirmation rate* approach yielded a similar estimate to those from the *MI* approach, but a standard error of 0.29 smaller than the 0.35 from the *MI* approach, and empirical power of 59.9%, which was greater than power of 48.8% from the *MI* approach.

Regardless of whether the PCR was the same or different for exposed and unexposed cases and regardless of the proportion of reviewed cases, the *MI* approach always produced an estimate for $\hat{\beta}_1$ that was comparable to those from the gold standard. As expected, the standard error of $\hat{\beta}_1$ increased and the empirical power decreased with fewer reviewed cases. In our simulations, the standard error of $\hat{\beta}_1$ from the *MI* approach was always smaller than or approximately equal to the ones from the *confirmed cases only* approach, but never smaller than the gold standard or the *known confirmation rate* approach.

Monte Caro errors (MCE) are also presented in Tables 3 and 4 for $\beta_0 = -12$ and $\beta_0 = -13$, respectively. Among the four approaches, the *observed cases* approach had the largest sample size because it used all observed cases, followed by the *known confirmation rate* approach and the *MI* approach that used both confirmed cases and those not chart reviewed, and the *confirmed cases only* approach had the smallest sample size because it used only confirmed cases. Consistent with sample sizes, across all scenarios, the *observed cases* approach yielded the smallest MCE, followed by the *known confirmation rate* approach and the *MI* approach, and the *confirmed cases only* approach yielded the largest MCE. For three approaches: the *observed cases* approach, the *confirmed cases only* approach and the *MI* approach, MCE was comparable to the mean standard error of $\hat{\beta}_1$, indicating that these three approaches produced unbiased standard error of $\hat{\beta}_1$. When the proportion of reviewed cases were low (i.e., $R_{M0} = R_{M1} = R_{L0} = R_{L1} = 50\%$), the *known confirmation rate* approach yielded MCE greater than the estimated standard error of $\hat{\beta}_1$ (e.g., 0.20 versus 0.17 for $R_0 = R_1 = 50\%$) when $\beta_0 = -12$, indicating that the *known confirmation rate* approach underestimated the standard error of $\hat{\beta}_1$. The difference was even greater when $\beta_0 = -13$ (e.g., 0.38 versus 0.29 for $R_0 = R_1 = 50\%$).

The trend of bias was similar when the adverse event was rare ($\beta_0 = -13$) except that the standard error of the vaccination effect was larger for $\beta_0 = -13$ and the *MI* approach was even more effective in reducing type I error and bias as compared to the other three approaches.

4 | AN EXAMPLE

A recent study evaluating the safety of live attenuated influenza vaccine (LAIV) in children ages 2–17 years old during the 2003–2004 through 2012–2013 influenza seasons used the proposed approaches from this paper in an analysis of syncopal events following vaccination (Daley et al. 2018). We used a SCCS design to examine adverse events following LAIV

vaccinations. The day a child received LAIV was considered as the one-day risk interval; all other time during the influenza season was considered the control interval. Six VSD sites participated in the study (Marshfield Clinic; Kaiser Permanente (KP) Northwest; KP Washington; KP Northern California; KP Southern California; and KP Colorado), and 532 presumptive cases of syncope were identified in the control interval (unexposed cases), and 11 presumptive cases were identified in the risk interval (exposed cases). All 11 exposed were reviewed with medical charts. Among the 532 unexposed cases, 440 were from two large sites, KP Northern California and KP Southern California. Only 96 (22%) of the 440 unexposed cases were reviewed from these two large sites, and all unexposed cases from smaller sites were reviewed. The PCRs were 45% and 69% for exposed and unexposed cases, respectively. Because exposed cases had lower PCR (i.e., more false exposed cases), we expected that the *observed cases* approach would overestimate IRR. We then analyzed the SCCS data using the four proposed approaches from this paper. The estimated IRRs (95% CIs) were 4.49 (2.47–8.16), 8.01 (3.28–19.56), 2.51 (1.04–6.05), and 2.52 (1.04–6.10), for the *observed cases, confirmed case only*, *known confirmation rate*, and *MI* approaches, respectively. Based on the results from our simulation study, we reported the estimated IRR from the *MI* approach since this approach yielded a proper type I error rate, a largely unbiased point estimate, a proper variance estimate, and appropriate statistical power (Daley et al., 2018). For the LAIV example, the number of exposed cases was far smaller than the number of unexposed cases (11 versus 532) and they were all reviewed. The number of exposed cases was the determining factor for the standard errors of coefficients; the uncertainty of PCRs of the unexposed was expected to have a limited impact on estimating the standard error of coefficients. Thus, the *known confirmation rate* and *MI* methods produced similar point estimates and confidence intervals.

5 | DISCUSSION

We conducted a simulation study to evaluate the performance of four approaches for analyzing SCCS data with misclassification of cases when confirmation rates are available from an internal validation sample. Clearly, analyzing *observed cases* alone resulted in biased estimates of the vaccination effect when the proportion of true observed cases differed between risk and control intervals. When the true PCR was equal in exposed and unexposed cases, the *observed cases* approach yielded unbiased estimates and greater statistical power. With this method, the use of misclassified cases increased the sample size. Although the increased sample size had no impact on type I error when PCR was the same for exposed and unexposed cases simply because the test was arranged to control the type I error (Table 2), it had impact on statistical power. Under the alternative hypothesis (IRR = 2.0), the *observed cases* approach yielded smaller standard error of the coefficient (Tables 3 and 4) and greater empirical power (Table 5) than the gold standard due to increased sample size. In addition, use of the misclassified cases (the *observed cases* approach) requires the assumption of equal PCRs. If the assumption is violated, the *observed cases* approach will yield biased estimates. PCRs will differ by exposure if either the underlying outcome specificity differs by exposure, or if there is a true exposure effect. Therefore, we do not recommend using the *observed cases* approach when misclassification of cases is present. The *confirmed cases* only approach produced unbiased estimates if the proportions of

observed cases that were medically chart reviewed were the same by exposure status, but it produced larger standard errors and thus lower empirical power compared to the gold standard and the *MI* approach. When all exposed cases are reviewed but only some unexposed cases are reviewed, such as in the LAIV safety study example, the *confirmed cases* only approach should never be considered because it is subject to large bias. The *known confirmation rate* approach may produce unbiased estimates, but the standard error of the vaccination effect is underestimated when the proportion of reviewed cases is low; thus, this method may lead to inflated empirical power. By contrast, regardless of whether the PCR is the same or different by exposure status and whether the proportion of observed cases chart reviewed differs by exposure status, the *multiple imputation* approach consistently yielded unbiased estimates of the vaccination effect and standard errors that increased as the number of charts reviewed decreased.

In a vaccine safety study, adjudicating all presumptive cases identified in EHR data and then including only confirmed cases in the final analyses is ideal (the gold standard in our simulation). When there is a relatively large number of presumptive cases and resources are limited, researchers may be able to limit the length of the control period, thus reducing the number of presumptive unexposed cases to review. Another option is to review only a proportion of cases, and we showed through simulation that to obtain an unbiased estimate using the *confirmed case only* approach, the proportion of reviewed cases must be the same for exposed and unexposed cases. However, compared to the *confirmed case only* approach, the inclusion of cases that are not medically chart reviewed can increase statistical power by using the *MI* approach regardless of how the internal validation sample is chosen. We emphasize the importance of using the *MI* approach when researchers chart review all exposed cases but only a proportion of unexposed cases. Other approaches can lead to a biased estimate and/or standard error of the vaccination effect. Consequently, it may lead to false conclusions about the association between vaccination and adverse events.

A key limitation to adjudicating all or a sample of presumptive cases identified in EHR data is that this approach presumes that outcome specificity may be $<100\%$, but assumes that outcome sensitivity is 100%. A primary reason for outcome false negatives in EHR data is differences in health care-seeking behavior (Chubak et al., 2012). However, with severe and acute adverse events, such as syncope or seizures in children, most people will rapidly seek care. Therefore, imperfect outcome sensitivity is likely more of a concern with less severe outcomes where there may be varying propensity for individuals to seek care.

There are several limitations to this study. First, age and seasonality are often associated with vaccination and adverse events in vaccine safety studies. We did not include covariates such as age and seasonality in the simulation. Second, we did not consider the potential misclassification of exposure. It is possible that some individuals' vaccination status could be misclassified if an incorrect vaccine or vaccination date is recorded in EHR data. Third, we did not consider false negative cases because they were not identifiable using EHR data. If false negative cases existed and the false negative rate differed between exposed and unexposed cases, this type of misclassification could bias estimation of a vaccination effect.

In conclusion, our simulation results suggest that the *observed cases, confirmed case only*, and *known confirmation rate* approaches may inflate the type I error, yield biased point estimates, and result in inflated or reduced statistical power when confirmation rates of cases are not perfect. The *multiple imputation* approach considers the uncertainty of estimated confirmation rates from an internal validation sample, yields acceptable type I error rates near 5%, produces largely unbiased point estimates and variance estimates, and yields proper statistical power. To address the influence of false positive cases in EHR data, researchers should consider using the *multiple imputation* approach when only a sample of presumptive cases can be validated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was funded by the Centers for Disease Control and Prevention (CDC) as part of the Vaccine Safety Datalink project (contract #200-2012-53582). Xu was also supported by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780.

Funding information

Centers for Disease Control and Prevention, Grant/Award Number: 200-2012-53582; NIH, Grant/Award Number: UL1 RR025780; NCRR

REFERENCES

Baggs J, Gee J, Lewis E, Fowler G, Benson P, Lieu T, … Weintraub E (2011). The Vaccine Safety Datalink: A model for monitoring immunization safety. Pediatrics, 127 (Suppl 1), S45–S53. [PubMed: 21502240]

Chubak J, Pocobelli G, & Weiss NS (2012). Tradeoffs between accuracy measures for electronic health care data algorithms. Journal of Clinical Epidemiology, 65 (3), 343–349. [PubMed: 22197520]

Cole SR, Chu H, & Greenland S (2006). Multiple-imputation for measurement-error correction. International Journal of Epidemiology, 35 (4), 1074–1081. [PubMed: 16709616]

Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, … Brown JS (2012). Design considerations, architecture, and use of the mini-sentinel distributed data system. Pharmacoepidemiology and Drug Safety, 21 (Suppl. 1), 23–31. [PubMed: 22262590]

Daley MF, Clarke CL, Glanz JM, Xu S, Hambidge SJ, Donahue JG, … Weintraub E (2018). The safety of live attenuated influenza vaccine in children and adolescents 2 through 17 years of age: A Vaccine Safety Datalink study. Pharmacoepidemiology and Drug Safety, 27 (1), 59–68. [PubMed: 29148124]

Edwards JK, Cole SR, Troester MA, & Richardson DB (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. American Journal of Epidemiology, 177 (9), 904–912. [PubMed: 24627573]

Farrington CP (1995). Relative incidence estimation from case series for vaccine safety evaluation. Biometrics, 51, 228–235. [PubMed: 7766778]

Glanz JM, Newcomer SR, Jackson ML, Omer SB, Bednarczyk RA, Shoup JA, … Daley MF (2016). White paper on studying the safety of the childhood immunization schedule in the Vaccine Safety Datalink. Vaccine, 34 (Suppl 1), A1–A29. [PubMed: 26830300]

Hambidge SJ, Glanz JM, France EK, McClure D, Xu S, Yamasaki K, … DeStefano F, & the Vaccine Safety Datalink Team. (2006). Safety of trivalent inactivated influenza vaccine in children 6 to 23 months old. Journal of the American Medical Association, 296 (16), 1990–1997. [PubMed: 17062862]

Hambidge SJ, Newcomer SR, Narwaney KJ, Glanz JM, Daley MF, Xu S, … DeStefano F (2014). Timely versus delayed early childhood vaccination and seizures. Pediatrics, 133, 1492–1499.

Hersh WR (2007). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. The American Journal of Managed Care, 13, 277–278. [PubMed: 17567224]

Koehler E, Brown E, & Haneuse SJ-PA (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. The American Statistician, 63, 155–162. [PubMed: 22544972]

Lyles RH, Tang L, Superak HM, King CC, Celentano DD, Lo Y, & Sobel JD (2011). Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. Epidemiology, 22 (4), 589–597. [PubMed: 21487295]

McNeil MM, Gee J, Weintraub ES, Belongia EA, Lee GM, Glanz JM, … DeStefano F (2014). The Vaccine Safety Datalink: Successes and challenges monitoring vaccine safety. Vaccine, 32 (42), 5390–5398. [PubMed: 25108215]

McNeil MM, Weintraub ES, Duffy J, Sukumaran L, Jacobsen SJ, Klein NP, … DeStefano F (2016). Risk of anaphylaxis after vaccination in children and adults. The Journal of Allergy and Clinical Immunology, 137 (3), 868–878. [PubMed: 26452420]

Miller E, Waight P, Farrington P, Stowe J, & Taylor B (2001). Idiopathic thrombocytopenic purpura and MMR vaccine. Archives of Disease in Childhood, 84, 227–229. [PubMed: 11207170]

Mullooly J, Drew L, DeStefano F, Maher J, Bohlke K, Immanuel V, … Chen R (2004). Quality assessments of HMO diagnosis databases used to monitor childhood vaccine safety. Methods of Information in Medicine, 43, 163–170. [PubMed: 15136866]

Mullooly JP, Donahue JG, DeStefano F, Baggs J, & Eriksen E, & VSD Data Quality Working Group. (2008). Predictive value of ICD-9-CM codes used in vaccine safety research. Methods of Information in Medicine, 47, 328–335. [PubMed: 18690366]

Nguyen M, Ball R, Midthun K, & Lieu TA (2012). The Food and Drug Administration's postlicensure rapid immunization safety monitoring program: Strengthening the federal vaccine safety enterprise. Pharmacoepidemiology and Drug Safety, 21 (Suppl. 1), 291–297. [PubMed: 22262619]

Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, & McClellan M (2009). The new Sentinel Network–improving the evidence of medical-product safety. The New England Journal of Medicine, 361, 645–647. [PubMed: 19635947]

Platt R, & Carnahan R (2012). (eds). The U.S. Food and Drug Administration's Mini-Sentinel Program: Status and direction. Pharmacoepidemiology and Drug Safety, 21 (S1), 1–8.

Rubin DB (1987). Multiple imputation for nonresponse in surveys New York, NY: Wiley.

Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE, & Expert Panel (2007). Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, 14 (1), 1–9. [PubMed: 17077452]

Stowe J, Andrews N, Ladhani S, & Miller E (2016). The risk of intussusception following monovalent rotavirus vaccination in England: A self-controlled case-series evaluation. Vaccine, 34, 3684–3689. [PubMed: 27286641]

Sun Y, Christensen J, Hviid A, Li J, Vedsted P, Olsen J, & Vestergaard M (2012). Risk of febrile seizures and epilepsy after vaccination with diphtheria, tetanus, acellular pertussis, inactivated poliovirus, and haemophilus influenzae type b. JAMA, 307 (8), 823–831. [PubMed: 22357833]

Weiskopf NG, & Weng C (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20, 144–151. [PubMed: 22733976]

Xu S, Newcomer S, Nelson J, Qian L, McClure D, Pan Y, … Glanz J (2014). Signal detection of adverse events with imperfect confirmation rates in vaccine safety studies using self-controlled case series design. Biometrical Journal, 56, 513–525. [PubMed: 24402780]

Xu S, Zhang L, Zeng C, Nelson J, Mullooly J, McClure D, & Glanz J (2011). Identifying optimal risk windows for self-controlled case series studies of vaccine safety. Statistics in Medicine, 30, 742–752. [PubMed: 21394750]

Xu S, Hambidge SJ, McClure DL, Daley MF, & Glanz JM (2013). A scan statistic for identifying optimal risk windows in vaccine safety studies using self-controlled case series design. Statistics in Medicine, 32, 3290–3299. [PubMed: 23303643]

TABLE 1

Number of observed cases when risk interval = 42 days and $\beta_1 = 0.69$

	Confirmation rate (%)		Number of observed cases					
β_0	Unexposed case (q_0)	Exposed case (q_1)	S_0	S_1	M_0	M_1	L_0	L_1
	50	50	33	9	100	27	201	53
−12	80	50	21	9	62	27	125	53
	50	80	33	6	100	17	201	33
	50	50	12	3	37	10	74	20
−13	80	50	8	3	23	10	46	20
	50	80	12	2	37	6	74	12

S_0, unexposed cases from the small site; S_1, exposed cases from the small site; M_0, unexposed cases from the medium site; M_1, exposed cases from the medium site; L_0, unexposed cases from the large site; L_1, exposed cases from the large site.

TABLE 2

Type I error rates when risk interval = 42 days, $\beta_1 = 0$ (incidence rate ratio = 1), and all cases at small sites were reviewed

Confirmation rates (%)		Proportion of reviewed cases (%)				Type I error rates (%)					
Unexposed (q_0)	Exposed (q_1)					Observed cases		Confirmed case only		Known confirmation rate		Multiple imputation	
		R_{M0}	R_{M1}	R_{L0}	R_{L1}	$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$
50	50	100	100	100	100	3.7	5.0	4.2[a]	3.5[a]	NA	NA	NA	NA
		80	80	80	80			3.0	2.9	5.2	4.4	3.9	3.4
		50	50	50	50			5.2	2.9	11.5	9.6	5.3	5.0
		80	100	80	100			16.3	8.0	4.9	3.9	4.8	4.3
		50	100	50	100			69.2	30.2	4.9	3.9	4.5	3.7
80	50	100	100	100	100	78.9	41.8	5.4[a]	3.4[a]	NA	NA	NA	NA
		80	80	80	80			4.8	2.7	4.7	4.3	3.8	4.0
		50	50	50	50			3.5	2.2	9.8	7.9	4.6	4.7
		80	100	80	100			17.0	8.2	4.9	2.8	5.2	3.1
		50	100	50	100			71.9	33.9	5.4	3.3	5.3	3.5
50	80	100	100	100	100	69.0	24.2	5.5[a]	3.5[a]	NA	NA	NA	NA
		80	80	80	80			3.6	2.9	4.9	4.4	4.3	3.6
		50	50	50	50			4.0	2.4	7.9	7.9	4.4	5.2
		80	100	80	100			14.4	10.1	4.1	4.5	4.2	4.8
		50	100	50	100			68.0	34.5	5.5	4.6	4.9	4.1

[a] Indicates gold standard of analysis of *confirmed cases only* when 100% of observed cases have been reviewed

R_{M0}, proportion of reviewed unexposed cases from the medium site; R_{M1}, proportion of reviewed exposed cases from the medium site; R_{L0}, proportion of reviewed unexposed cases from the large site; R_{L1}, proportion of reviewed exposed cases from the large site.

TABLE 3

Mean $\hat{\beta}_1$ (mean standard error of $\hat{\beta}_1$), Monte Carlo error (MCE) of $\hat{\beta}_1$, and percent bias when risk interval = 42 days, $\beta_1 = 0.69$ (incidence rate ratio = 2), $\beta_0 = -12$, and all cases at small sites were reviewed

Confirmation rates (%)		Proportion of reviewed cases (%)				Mean (standard error) of $\hat{\beta}_1$ MCE, percent bias			
Unexposed (q_0)	**Exposed (q_1)**	**R_{M0}**	**R_{M1}**	**R_{L0}**	**R_{L1}**	**Observed cases**	**Confirmed case only**	**Known confirmation rate**	**Multiple imputation**
50	50	100	100	100	100	0.69 (0.12), 0.11, 0%	0.68 (0.17), 0.17, −1%[a]	NA	NA
		80	80	80	80		0.69 (0.19), 0.18, 0%	0.69 (0.17), 0.17, 0%	0.69 (0.18), 0.17, 0%
		50	50	50	50		0.70 (0.23), 0.24, 1%	0.69 (0.17), 0.20, 0%	0.68 (0.20), 0.20, −1%
		80	100	80	100		0.89 (0.17), 0.16, 30%	0.69 (0.17), 0.16, 0%	0.69 (0.17), 0.16, 0%
		50	100	50	100		1.28 (0.18), 0.18, 86%	0.68 (0.17), 0.17, −1%	0.68 (0.18), 0.17, −1%
80	50	100	100	100	100	1.16 (0.13), 0.12, 68%	0.69 (0.17), 0.16, 0%[a]	NA	NA
		80	80	80	80		0.69 (0.19), 0.18, 0%	0.69 (0.17), 0.17, 0%	0.69 (0.18), 0.17, 0%
		50	50	50	50		0.68 (0.23), 0.23, −1%	0.68 (0.17), 0.20, −1%	0.68 (0.20), 0.20, −1%
		80	100	80	100		0.89 (0.17), 0.16, 30%	0.69 (0.17), 0.16, 0%	0.70 (0.17), 0.16, 1%
		50	100	50	100		1.29 (0.18), 0.18, 87%	0.69 (0.17), 0.17, 0%	0.69 (0.17), 0.17, 0%
50	80	100	100	100	100	0.22 (0.15), 0.14, −68%	0.69 (0.17), 0.17, 0%[a]	NA	NA
		80	80	80	80		0.69 (0.19), 0.19, 0%	0.69 (0.17), 0.18, 0%	0.69 (0.18), 0.18, 0%
		50	50	50	50		0.68 (0.23), 0.23, −1%	0.69 (0.17), 0.19, 0%	0.68 (0.19), 0.19, −1%
		80	100	80	100		0.89 (0.17), 0.17, 30%	0.69 (0.17), 0.17, 0%	0.69 (0.17), 0.17, 0%
		50	100	50	100		1.28 (0.18), 0.18, 86%	0.69 (0.17), 0.17, 0%	0.68 (0.18), 0.17, −1%

[a]Indicates gold standard of analysis of *confirmed cases only* when 100% of observed cases have been reviewed

R_{M0}, proportion of reviewed unexposed cases from the medium site; R_{M1}, proportion of reviewed exposed cases from the medium site; R_{L0}, proportion of reviewed unexposed cases from the large site; R_{L1}, proportion of reviewed exposed cases from the large site.

TABLE 4

Mean $\hat{\beta}_1$ (mean standard error of $\hat{\beta}_1$), Monte Carlo error (MCE) of $\hat{\beta}_1$, and percent bias when risk interval = 42 days, $\beta_1 = 0.693$ (incidence rate ratio = 2), $\beta_0 = -13$, and all cases at small sites were reviewed

Confirmation rates (%)		Proportion of reviewed cases (%)				Mean (standard error) of $\hat{\beta}_1$ MCE, percent bias			
Unexposed (q_0)	**Exposed (q_1)**	R_{M0}	R_{M1}	R_{L0}	R_{L1}	**Observed cases**	**Confirmed case only**	**Known confirmation rate**	**Multiple imputation**
50	50	100	100	100	100	0.68 (0.20), 0.20, −1%	0.66 (0.29), 0.29, −4%[a]	NA	NA
		80	80	80	80		0.67 (0.32), 0.31, −3%	0.66 (0.30), 0.30, −4%	0.66 (0.30), 0.30, −4%
		50	50	50	50		0.63 (0.40), 0.41, −9%	0.63 (0.29), 0.38, −9%	0.62 (0.35), 0.37, −10%
		80	100	80	100		0.86 (0.29), 0.29, 25%	0.66 (0.29), 0.29, −4%	0.66 (0.29), 0.29, −3%
		50	100	50	100		1.28 (0.31), 0.31, 86%	0.68 (0.28), 0.30, −1%	0.68 (0.30), 0.30, −1%
80	50	100	100	100	100	1.16 (0.21), 0.21, 68%	0.68 (0.29), 0.29, −1%[a]	NA	NA
		80	80	80	80		0.67 (0.32), 0.33, −3%	0.68 (0.28), 0.31, −1%	0.67 (0.30), 0.31, −3%
		50	50	50	50		0.65 (0.39), 0.42, −6%	0.65 (0.29), 0.37, −6%	0.64 (0.34), 0.37, −7%
		80	100	80	100		0.86 (0.29), 0.30, 25%	0.67 (0.29), 0.29, −3%	0.67 (0.29), 0.29, −3%
		50	100	50	100		1.27 (0.31), 0.30, 84%	0.67 (0.29), 0.29, −3%	0.68 (0.29), 0.29, −1%
50	80	100	100	100	100	0.20 (0.24), 0.24, −71%	0.67 (0.28), 0.29, −3%[a]	NA	NA
		80	80	80	80		0.67 (0.32), 0.32, −3%	0.67 (0.28), 0.30, −3%	0.66 (0.29), 0.30, −3%
		50	50	50	50		0.64 (0.39), 0.41, −7%	0.65 (0.29), 0.36, −6%	0.65 (0.33), 0.36, −6%
		80	100	80	100		0.88 (0.29), 0.29, 28%	0.68 (0.28), 0.29, −1%	0.69 (0.29), 0.30, 0%
		50	100	50	100		1.28 (0.31), 0.32, 86%	0.68 (0.28), 0.30, −1%	0.69 (0.30), 0.30, 0%

[a]Indicates gold standard of analysis of *confirmed cases only* when 100% of observed cases have been reviewed

R_{M0}, proportion of reviewed unexposed cases from the medium site; R_{M1}, proportion of reviewed exposed cases from the medium site; R_{L0}, proportion of reviewed unexposed cases from the large site; R_{L1}, proportion of reviewed exposed cases from the large site.

TABLE 5

Empirical power when risk interval = 42 days, $\beta_1 = 0.69$ (incidence rate ratio = 2), and all cases at small sites were reviewed

Confirmation rates (%)		Proportion of reviewed cases (%)				Empirical power (%)							
Unexposed (q_0)	Exposed (q_1)	R_{M0}	R_{M1}	R_{L0}	R_{L1}	Observed cases		Confirmed case only		Known confirmation rate		Multiple imputation	
						$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$	$\beta_0 = -12$	$\beta_0 = -13$
50	50	100	100	100	100	100	90.7	95.9[a]	64.3[a]	NA	NA	NA	NA
		80	80	80	80			93.7	56.8	96.4	63.2	95.5	60.0
		50	50	50	50			81.4	39.6	94.0	59.9	89.1	48.8
		80	100	80	100			99.7	80.2	97.3	61.6	97.0	62.1
		50	100	50	100			100	95.3	96.9	65.3	96.3	63.9
80	50	100	100	100	100	100	99.8	97.0[a]	65.9[a]	NA	NA	NA	NA
		80	80	80	80			93.7	58.3	95.8	65.4	94.9	62.8
		50	50	50	50			81.0	42.5	93.3	63.0	88.4	54.9
		80	100	80	100			99.8	80.6	97.6	66.1	97.6	67.2
		50	100	50	100			100	95.9	96.2	63.6	96.1	64.6
50	80	100	100	100	100	35.8	16.3	96.6[a]	64.9[a]	NA	NA	NA	NA
		80	80	80	80			92.4	57.8	95.7	63.3	94.7	62.0
		50	50	50	50			81.5	42.2	95.4	63.0	92.2	59.0
		80	100	80	100			99.6	82.7	96.0	65.6	96.2	67.2
		50	100	50	100			100	96.1	95.9	65.4	95.4	63.8

[a]Indicates gold standard of analysis of *confirmed cases only* when 100% of observed cases have been reviewed

R_{M0}, proportion of reviewed unexposed cases from the medium site; R_{M1}, proportion of reviewed exposed cases from the medium site; R_{L0}, proportion of reviewed unexposed cases from the large site; R_{L1}, proportion of reviewed exposed cases from the large site.