Received: 23 May 2023 | Revised: 3 November 2023 | Accepted: 8 November 2023

DOI: 10.1002/sim.9968

RESEARCH ARTICLE

Statistics in Medicine WILEY

Bayesian safety surveillance with adaptive bias correction

Fan Bu[1,2] | Martijn J. Schuemie[1,3] | Akihiko Nishimura[4] | Louisa H. Smith[5,6] | Kristin Kostka[6] | Thomas Falconer[7] | Jody-Ann McLeggon[7] | Patrick B. Ryan[3] | George Hripcsak[7] | Marc A. Suchard[1]

[1]Department of Biostatistics, University of California, Los Angeles, California, USA

[2]Department of Biostatistics, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

[3]Janssen Research and Development, Raritan, New Jersey, USA

[4]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA

[5]Department of Health Sciences, Northeastern University, Portland, Maine, USA

[6]The OHDSI Center at the Roux Institute, Northeastern University, Portland, Maine, USA

[7]Department of Biomedical Informatics, Columbia University, New York, New York, USA

Correspondence

Fan Bu, Department of Biostatistics, University of California, Los Angeles, CA, USA.
Email: fanbu42@gmail.com

Funding information

NIH Common Fund, Grant/Award Numbers: R01 AG068002, R01 LM006910

Postmarket safety surveillance is an integral part of mass vaccination programs. Typically relying on sequential analysis of real-world health data as they accrue, safety surveillance is challenged by sequential multiple testing and by biases induced by residual confounding in observational data. The current standard approach based on the maximized sequential probability ratio test (MaxSPRT) fails to satisfactorily address these practical challenges and it remains a rigid framework that requires prespecification of the surveillance schedule. We develop an alternative Bayesian surveillance procedure that addresses both aforementioned challenges using a more flexible framework. To mitigate bias, we jointly analyze a large set of negative control outcomes that are adverse events with no known association with the vaccines in order to inform an empirical bias distribution, which we then incorporate into estimating the effect of vaccine exposure on the adverse event of interest through a Bayesian hierarchical model. To address multiple testing and improve on flexibility, at each analysis timepoint, we update a posterior probability in favor of the alternative hypothesis that vaccination induces higher risks of adverse events, and then use it for sequential detection of safety signals. Through an empirical evaluation using six US observational healthcare databases covering more than 360 million patients, we benchmark the proposed procedure against MaxSPRT on testing errors and estimation accuracy, under two epidemiological designs, the historical comparator and the self-controlled case series. We demonstrate that our procedure substantially reduces Type 1 error rates, maintains high statistical power and fast signal detection, and provides considerably more accurate estimation than MaxSPRT. Given the extensiveness of the empirical study which yields more than 7 million sets of results, we present all results in a public R ShinyApp. As an effort to promote open science, we provide full implementation of our method in the open-source R package `EvidenceSynthesis`.

KEYWORDS

Bayesian sequential testing, postmarket safety surveillance, real-world evidence, systematic error

1 | INTRODUCTION

Mass vaccination is a cornerstone of effective disease control.[1-4] Addressing any safety concerns over new vaccine products is therefore essential. Despite preapproval clinical trials that ensure vaccine effectiveness and safety, rare but severe adverse events often go undetected due to limited sample sizes and statistical power. Postapproval safety surveillance is thus a critical component of mass vaccination programs. One common approach of safety surveillance is to sequentially analyze real-world observational data as they accrue over time.[5-15]

However, there are two major challenges in safety surveillance using observational data. One challenge is to adjust for sequential multiple testing with a flexible surveillance schedule while allowing for rapid detection of true safety signals. Another challenge is to account for and correct for residual systematic error in the observational data that induces bias and can inflate decision errors.[16-18] The maximized sequential probability ratio test (MaxSPRT) stands as a standard and commonly adopted statistical framework for sequential safety surveillance.[19] MaxSPRT, however, does not provide a fully satisfactory solution to these challenges.[20,21] MaxSPRT remains a rigid framework that requires a prefixed surveillance schedule and does not allow the analysis plan to adapt to real-world data accrual behaviors. Morevoer, MaxSPRT does not naturally allow for extensions to correct for biases in observational studies.

In this work, we propose an alternative safety surveillance procedure that addresses these two challenges in a unified and interpretable framework. We develop the procedure based on Bayesian sequential analysis, with an empirical modeling component that adaptively corrects for biases. More specifically, at each sequential analysis timepoint, we perform Bayesian inference to obtain a posterior distribution for the effect of vaccination on the adverse event of interest; this is augmented by an adaptive bias correction approach that builds on an empirical bias distribution learned through analysis of negative control outcomes. From the posterior distribution, we compute the posterior probability of the hypothesis that vaccination increases the risk of the adverse event, given the data accrued. We then use the posterior probability as a test statistic for signal detection: at the first timepoint when the posterior probability exceeds a user-specified threshold, we declare a safety signal. Such surveillance procedure, unlike MaxSPRT, does not require a prespecified surveillance schedule, but instead only relies on data evidence that has already accrued. We will detail our proposed framework in Section 3.

Through comprehensive empirical evaluations using large-scale observational healthcare databases of more than 360 million unique patients, we demonstrate the benefits of our methods. Compared to the standard approach, our Bayesian surveillance procedure offers greater flexibility in the surveillance schedule, improved transparency and interpretability for evidence extraction and decision-making, and more reliable error control through bias correction. In the remainder of this section, we provide necessary background information about safety surveillance as a statistical problem.

1.1 | Safety surveillance as a sequential hypothesis test

In safety surveillance, the key question we wish to answer is: for a particular vaccine (eg, seasonal flu vaccine) and an adverse event outcome (eg, the Guillain-Barré Syndrome), does taking the vaccine increase the incidence rate of this adverse event compared to unvaccinated?

The key quantity of interest to estimate in order to answer such question is the incidence rate ratio, *RR*, defined as the ratio between the population incidence rate during a risk window post vaccination versus that without vaccination or during a control window. Thus, safety surveillance is essentially a hypothesis testing problem regarding the value of *RR*:

$$H_0 : RR \leq 1 \qquad \text{vs.} \qquad H_1 : RR > 1. \tag{1}$$

Here, the alternative hypothesis H_1 indicates an increased risk of the adverse event due to vaccine exposure and thus, rejecting H_0 in favor H_1 implies raising a safety concern (or a safety signal).

Oftentimes we perform inference on the natural logarithm of *RR*, that is, with quantity of interest $\beta = \log(RR)$. Then the equivalent hypotheses are:

$$H_0 : \beta \leq 0 \qquad \text{vs.} \qquad H_1 : \beta > 0. \tag{2}$$

In the context of safety surveillance, we usually rely on real-world observational data that are updated *sequentially over time*. Example data sources include administrative claims, electronic health records (EHRs), local registries and spontaneous reports. In this article, we mainly utilize claims and EHRs. In order to rapidly detect safety signals, we need to perform *sequential* hypothesis test at discrete time points when new batches of data have accrued. Suppose $T_{\max}$ denotes the maximum length of surveillance window and we perform analysis at time $t = 1, 2, \dots, T_{\max}$ ($T_{\max}$ can be infinite if a maximum length is not preset). Let $\mathbf{X}_t$ represent *all* data that have accrued up to t since the beginning of surveillance. At each timepoint t, we base analysis on $\mathbf{X}_t$: as soon as the hypothesis test informed by $\mathbf{X}_t$ leads us to reject H_0 in favor of H_1, we stop the surveillance study and declare a safety signal at time t.

There are two essential elements in sequential testing: first, a test statistic $W_t(\mathbf{X}_t)$ computed given $\mathbf{X}_t$ at each timepoint t; second, a decision rule that decides if we reject H_0 at time t, usually via checking if $W_t > A_t$ for a prespecified threshold A_t. Different statistical frameworks differ on the definition and computation of W_t and A_t. We first briefly discuss the current standard approach MaxSPRT and then overview a Bayesian framework in these important aspects.

1.2 | MaxSPRT: Current standard approach for safety surveillance

MaxSPRT is a sequential testing approach developed by Reference 19 to account for sequential multiplicity due to repeated data analyses in safety surveillance. At each analysis time point t, the MaxSPRT test statistic is defined as the generalized log likelihood ratio (LLR) between the alternative hypothesis H_1 and null hypothesis H_0:

$$W_t(\mathbf{X}_t) := \log(\max \mathrm{LR}_t) = \log\left(\frac{\max_{H_1:\beta>0} p(\mathbf{X}_t|\beta)}{\max_{H_0:\beta\leq 0} p(\mathbf{X}_t|\beta)}\right), \tag{3}$$

where $p(\mathbf{X}_t|\beta)$ denotes the joint probability density function of data $\mathbf{X}_t$ (used as the likelihood function in regards to β) specified by the model of choice. In Reference 19, examples of data models include binomial and Poisson distributions for incidence counts of adverse events.

To calculate a decision threshold A_t, MaxSPRT adopts an alpha-spending approach. First, the user must prespecify the total length of surveillance (total number of analyses) and the incremental sample sizes between every two consecutive analyses. Then, given a desired significance level α (usually $\alpha = 0.05$), MaxSPRT numerically solves for a constant critical value cv under the assumed data model such that the total quota of α can be reasonably spent across all the analyses under the prespecified schedule to ensure that the total Type 1 error rate is bounded below α (see Reference 19 and the R package `sequential` for technical details). Finally, with threshold $A_t \equiv cv$, as soon as W_t exceeds cv, MaxSPRT rejects H_0 in favor of H_1 and declares a safety signal.

1.3 | Overview of Bayesian inference and sequential analysis

In a Bayesian framework, inference is based on posterior updates of beliefs from prior beliefs given observed data, which is naturally sequential as we can update beliefs whenever new data are observed.

Let $\pi_0(\beta)$ denote the *prior* probability density function for β, the parameter of interest. Given $p(\mathbf{X}_t|\beta)$(the likelihood function), we can obtain the *posterior* distribution $\pi_t(\beta|\mathbf{X}_t)$ using the Bayes rule:

$$\pi_t(\beta|\mathbf{X}_t) \propto p(\mathbf{X}_t|\beta)\pi_0(\beta). \tag{4}$$

Here, the symbol "∝" indicates that the left hand side is proportional to the right hand side by a factor of constant quantities.

One important feature of Bayesian hypothesis testing is that we can specify our *prior beliefs* on the two hypotheses H_0 and H_1 by assigning them with prior probabilities $P(H_0)$ and $P(H_1) = 1 - P(H_0)$. Arguably, $P(H_0)$ and $P(H_1)$ can be *any* arbitrary probabilities that sum to 1, which can be informed by prior knowledge about the hypotheses or the prior density function $\pi_0(\beta)$. For example, one convenient and common choice is to assign equal prior probabilities such that $P(H_0) = P(H_1) = \frac{1}{2}$.

In Bayesian sequential testing, the test statistic $W_t(\mathbf{X}_t)$ commonly takes two forms:

The first is the Bayes Factor, which accounts for the ratio of marginal data evidence in support of each hypothesis:

$$BF_{10}^{(t)} := \frac{m_1(\mathbf{X}_t)}{m_0(\mathbf{X}_t)},$$

where $m_i(\mathbf{X}_t) = \int_{\beta \in H_i} p_t(\mathbf{X}_t|\beta)\pi_0(\beta)d\beta$, for $i = 0, 1$. Common choices for the decision threshold A_t include 10, 20 and 30.[22-28] For example, using the Bayes Factor and taking 20 as the threshold, as soon as $BF_{10}^{(t)} > 20$, we stop the study and reject H_0 in favor of H_1.

The second is the posterior probability of either hypothesis. A straightforward test statistic $W_t(\mathbf{X}_t)$ for accepting H_1 is its posterior probability $P_{1,t}$ given the data $\mathbf{X}_t$ accrued up to time t:

$$W_t(\mathbf{X}_t) := P_{1,t} = P_t(H_1 : \beta > 0|\mathbf{X}_t) = \int_{H_1:\beta>0} \pi_t(\beta|\mathbf{X}_t)d\beta. \tag{5}$$

The decision threshold A_t for $P_{1,t}$ can take multiple values between 0 and 1. Common choices include 0.8, 0.9, and 0.95.[29-34]

1.3.1 | Organization of subsequent sections

The rest of this manuscript is structured as follows. In Section 2 we discuss the limitations of the standard approach for safety surveillance and motivate the development of an alternative framework. We propose our Bayesian procedure in Section 3 and then describe a large-scale empirical evaluation to benchmark its performance against MaxSPRT. We summarize results of the empirical evaluation in Section 4. We then briefly investigate the association between varicella zoster (Shingrix) vaccination and occurrences of the Guillain-Barré syndrome in Section 5 and finally conclude with discussions in Section 6.

2 | LIMITATIONS OF EXISTING APPROACHES

In this section, we illustrate the limitations of the existing safety surveillance framework. Through simple examples via simulations, we wish to show that (1) the standard MaxSPRT framework is inflexible and can lead to inconsistent decision-making, and (2) residual systematic error in data can bias analyses and requires additional correction.

Suppose we wish to learn the effect of a hypothetical vaccine exposure on some hypothetical adverse event outcome. We first assume a null outcome that has no relation to the exposure ($RR = 1$) to empirically evaluate Type 1 error rates over sequential analyses of data. Then we consider a positive outcome for which $RR = 2$ (ie, the hypothetical vaccine elevates the incidence rate by 2 fold) to examine the accuracy of estimating RR using two common epidemiological designs.

2.1 | MaxSPRT is inflexible and can produce inconsistent decisions

A key shortcoming of the MaxSPRT framework lies in its inflexibility—it requires users to prespecify the entire surveillance schedule, including the total length of the study, and the incremental sample sizes between analysis timepoints. This essentially requires predicting the data accrual timeline for at least a year into the future and committing to this prediction in order to design the sequential analysis plan. If reality deviates from the prediction (which often happens), then the analyst has to violate the prespecified plan or adopt some ad hoc plan, which can produce inconsistent or erroneous decisions.

We showcase this through a hypothetical simulation experiment where MaxSPRT is implemented on the same real-world dataset, but the analysis is performed by three analysts who have specified different analysis plans. For simplicity, suppose all three plans analyze data monthly and expect approximately 10 adverse event incidents during each month. However, they differ on the predicted total length of the surveillance: analyst **A** expects 24 months of data accrual and thus $10 \times 24 = 240$ total events will be expected at the end, analyst **B** expects 12 months and thus 120 total events, whereas analyst **C** expects 36 months and thus 360 total events. *Prior to seeing the real data*, each of the analyst precalculates their decision thresholds based on their prespecified analysis plan.

We simulate data by generating outcome incidence counts from a Poisson count model sequentially such that: (1) between two consecutive months, 10 events for the outcome are expected during the time interval; and (2) by the end of the data accrual process, $10 \times 24 = 240$ total events are expected, which equates to 24 months of data by expectation.

Upon seeing the real data that can disagree with the prespecified plans, the three analysts will end up carrying out different analyses:

A **"Oracle"**: analyst **A** correctly predicts the total number of data looks (24) which happen to agree with the actual length of data accrual.

B **"Hacky extension"**: analyst **B** only preplanned for **12** total data looks; however, data accrual extends beyond her original analysis endpoint, so in practice, if no signal is detected within 12 data looks, she may end up extending the surveillance to 24 analysis periods, which will increase Type 1 error due to over-spending the alpha.

C **"Early stop"**: analyst **C** preplanned **36** total data looks but the real data are only available up to 24 data looks. In this case, she has to stop the surveillance study early without spending all the planned alpha, which will likely decrease power.

Since the three analysts have adopted different surveillance plans (in terms of the total length), they would use *different* decision thresholds for MaxSPRT. Notably, the analysts cannot change the precalculated thresholds even if the real data deviate from their prespecified plans. Naturally, different decision thresholds will lead to *different* decisions made on the *same* simulated dataset, some of which are prone to increased errors.

In Figure 1 we plot the empirical Type 1 error rates accumulated over sequential analyses using the three different prespecified analysis plans. The Type 1 error rate is measured by the fraction of "reject H_0" decisions across 500 repeated simulations, where MaxSPRT is implemented with a prespecified significance level $\alpha = 5\%$.

The differences between the three curves reflect the difficulty of planning ahead in real-world studies. Only the "oracle" (analyst **A**) that predicts data accural length correctly and preplans surveillance accordingly is able to guarantee a close-to-nominal level of Type 1 error rate. The "early stop" plan (by analyst **C**) effectively does not exhaust the amount of "α" that is preplanned to spend over 36 data looks, and thus at 24 total data looks it would be under-powered. The "hacky extension" plan (by analyst **B**), on the other hand, overshoots on Type 1 error because it has preplanned for 12 analyses only (note that its Type 1 error rate is controlled at around 5% by 12 data looks). However, under the MaxSPRT framework (or a similar alpha-spending approach), if data behavior deviates from the a priori prediction (which happens often), there is no feasible option to adaptively adjust the analysis plan that allows us to use more available data or preserve power. Therefore, it is critical to develop a flexible framework that does *not* depend on prespecification of the surveillance schedule, which we will discuss in Section 3.1.

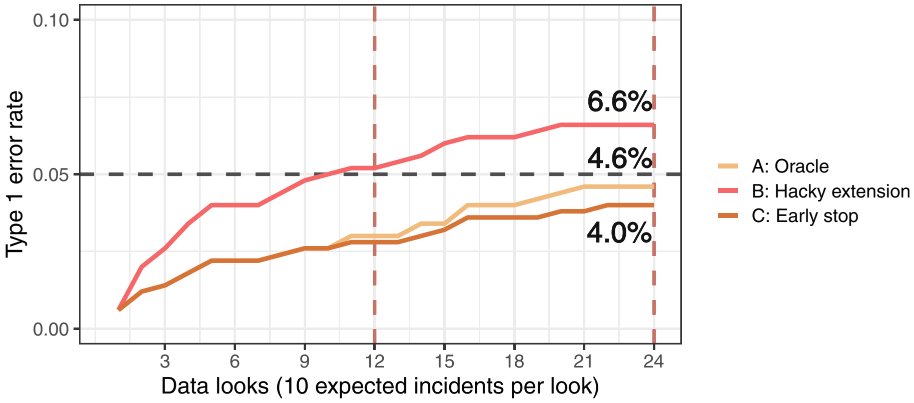


FIGURE 1 Inconsistent decisions by MaxSPRT due to required prespecified analysis plans. Empirical Type 1 error rate over 24 monthly surveillance data looks, with **same** data accrued monthly, only with different prespecified analysis plans. Error rates calculated for 500 simulations in which the synthesized exposure has **no effect** on the synthesized outcome. The dark gray horizontal line indicates the prespecified significance level, $\alpha = 5\%$. The end-of-analysis Type 1 error rates (by the end of 24 data looks) are annotated by text.

2.2 | Residual systematic error can bias analyses

We next present a toy simulated example where residual systematic error leads to biased estimation of RR. Residual systematic error is comprised of patterns, trends and covariates in the observational data that are not observed or measured, or not adjusted for in the study design or statistical model, but remain common because the data were collected for other uses than the study at hand.

We consider a safety surveillance study that lasts one year with data analysis scheduled at the end of each month (thus 12 total data looks); one year of historical data collected prior to the surveillance study are also available. Suppose the hypothetical exposure has a positive effect on a hypothetical outcome with $RR = 2$. We use two very common epidemiological designs for vaccine safety surveillance: the historical comparator[35-37] and the self-controlled case series (SCCS,[38-40]), to estimate RR through sequential analysis and examine the estimation accuracy.

We simulate subject-level trajectories of vaccine exposure and incidents of the adverse event outcome. In the simulations, we also inject common observational data features such as individual-level, temporal and seasonal effects that could result in bias with designs that typically fail to capture these features. More specifically, we simulate subject-level weekly incidence counts for the outcome of interest using the following Poisson count model:

$$\begin{aligned} &Y_{ik} \sim \text{Poisson}(\lambda_{ik}), \text{ where} \\ &\lambda_{ik} = \exp[\alpha_k + x_i\gamma + \mathbb{1}(i \text{ at risk in week } k) \times \beta]. \end{aligned}$$

Here Y_{ik} is the outcome incidence count for subject i during week k, where $k = 1, 2, \dots 104$ for which weeks 1 to 52 constitute the one-year historical period and weeks 53 to 104 constitute the present-time surveillance period. Further, x_i is a randomly assigned binary covariate that represents individual i's important characteristics, with coefficient γ. Intercept α_k denotes the expected weekly incidence rate without exposure, where there is a seasonality pattern as depicted in Figure 2A. The term $\mathbb{1}(i \text{ at risk in week } k)$ is a binary indicator that takes value 1 if subject i is inside the risk window post vaccination, and 0 otherwise. Recall that $\beta = \log(RR)$, with true value $\log(2)$ in this simulation. There is also a difference between the background incidence rates in the historical period (weeks 1 to 52) and in the surveillance period (weeks 53-104) where the average historical incidence rate captured in the data is only about 50% of the present-time rate. Such differential background rates between historical and present times are commonly seen in real-world data, due to factors such as data collection bias, misclassification, time-varying population effects, shifts in diagnostics standards and guidelines and so on.

Subject-level event trajectories (including a one-shot vaccine exposure and possibly multiple occurrences of the outcome) are simulated for $N = 5000$ individuals. We assume that each individual will be at risk of experiencing the

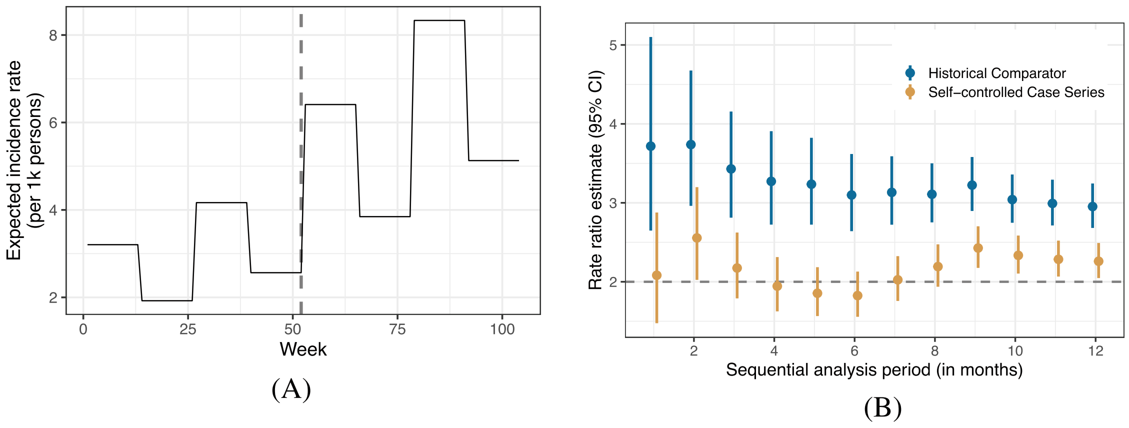


FIGURE 2 Two common epidemiological designs produce biased estimates due to residual systematic error. Settings and results from a simulation experiment with $N = 5000$ subject-level trajectories. (A) Time-varying background incidence rates of a hypothetical adverse event. Observations from the first year (week 1 to 52, left to gray vertical line) are used to calculate historical reference rates for the Historical Comparator design. (B) Rate ratio estimates (and 95% confidence intervals) obtained in monthly sequential analyses using two common epidemiological designs. The ground truth is $RR = 2$, but both methods produce biased estimates.

adverse event within the first 6 weeks following vaccine exposure (ie, 1-42 days post vaccination as the risk window). The vaccine uptake rate (number of people exposed to the hypothetical vaccine in each week) during week 53-104 has a seasonal pattern that takes the same shape as depicted in Figure 2A.

We implement the two epidemiological designs (and thus their implied statistical models) to construct counterfactual person-times that are compared to exposed person-times in order to estimate *RR*. Each design makes different assumptions about the subject-level weekly incidence rate λ_{ik} that could lead to bias induced by residual systematic error. Below, we briefly discuss each design and elaborate on the sources of bias later with the numerical results:

1. Historical comparator

This design constructs counterfactual person-times on unexposed historical individuals and compares present-time incidence rates of the outcome against historical reference rates from the counterfactuals to estimate *RR*. Here, we assume a Poisson regression model and take into account the seasonality effects within each year. That is,

$$\lambda_{ik} = \exp[s_k + x_i\gamma + \mathbb{1}(i \text{ at risk in week } k) \times \beta],$$

where s_k denotes the *week of the year* for week k; for example, both week 2 and week 54 will have $s_k = 2$, and thus week 2 would serve as the historical reference period for week 54. Notably, this design typically assumes that historical incidence rates are comparable to rates in present-time, which could induce bias.

2. SCCS

This design only considers "cases", that is, exposed subjects with at least one outcome event during the surveillance period. *Within each subject's individual trajectory*, SCCS compares outcome incidence during their risk window with that outside the risk window to estimate *RR*. Here, we implement a conditional Poisson regression model,

$$\lambda_{ik} = \exp[\gamma_i + \mathbb{1}(i \text{ at risk in week } k) \times \beta],$$

where γ_i represents an individual factor specific to person i, thus effectively modeling each subject-level trajectory as an inhomogeneous Poisson process. For SCCS, only present-time data (weeks 53-104) are used. We note that this design could be subject to bias with the existence of complex temporal effects such as seasonality.

Analysis using each design is performed sequentially every month (ie, every 4 or 5 weeks), on all data accrued up to the end of each month. For simplicity, we assume that there is no data accrual delay, such that events occurred in week k are observed promptly and available for analysis. That is, the third analysis is run on *all* data accumulated by the end of month 3.

Figure 2B presents *RR* estimates with 95% confidence intervals obtained by each analysis design at each analysis timepoint (by each month). Here, we estimate β as the parameter and then transform with $RR = \exp(\beta)$ to obtain point estimates and confidence intervals. The ground truth $RR = 2$ is marked by the gray horizontal dashed line. Neither design produces satisfactorily accurate estimates. The historical comparator severely over-estimates *RR*, as it fails to capture the time-varying effect in that the background incidence rate during the surveillance period is higher than that during the historical period. Such temporal trends in observational data that are confounded with vaccine exposures during surveillance are challenging to measure or model without knowing the true underlying data generative process. SCCS performs better, but its estimates can still be biased despite more data accrued over time. Even though SCCS can adjust for subject-level time-invariant covariates,[38,40] when the temporal patterns of vaccine exposures and outcome occurrences (and/or data accrual) are confounded, SCCS struggles to converge to the correct answer even with more data analyzed.

The complexity of residual systematic error in observational data goes far beyond the unmeasured temporal effects or confounders simulated here to illustrate how these commonly applied epidemiological designs are subject to bias. The true mechanisms of systematic error may be a combination of selection bias, misclassification, unmeasured confounding and many other factors that are not directly observable, testable or adjustable through an epidemiological design alone. This motivates us to consider an approach that can diagnose and correct for bias induced by residual systematic error, within the context of sequential analysis where bias should be adjusted for adaptively over time.

3 | METHODS

In this section, we first provide a description of our statistical method for sequential analysis under a Bayesian framework, with joint statistical modeling to adaptively correct for bias induced by residual systematic error. Then we outline the design of a comprehensive empirical evaluation of the proposed method on multiple large-scale administrative health databases. The full protocol of the empirical evaluation study is publicly available at https://suchard-group.github.io/Better/Protocol.html.

3.1 | Bayesian sequential analysis with bias correction

We adopt a joint statistical modeling approach to sequentially estimate β and adaptively correct for estimation bias induced by systematic error in observational data. The bias correction component of our model relies on estimating an empirical probabilistic distribution of the bias, accomplished with a fully data-driven approach.

3.1.1 | Overview of the Bayesian sequential analysis framework

The input of our method includes the data $\mathbf{X}_t$ accumulated up to each analysis timepoint t, a working statistical model that allows us to write down the joint density function $p(\mathbf{X}_t|\beta)$ (as the likelihood function), and a *prior distribution* $\pi_0(\beta)$ for β. We adopt a prior distribution such that the prior probabilities for H_0 and H_1 are both equal to $1/2$; one example prior choice to achieve this is to set $\pi_0(\beta)$ as a normal distribution with prior mean 0.

The output at each analysis time point t is a *posterior distribution* $\pi_t(\beta|\mathbf{X}_t)$ for β, obtained using the Bayes rule as in (4). In our framework we use the posterior probability for H_1, $P_{1,t}$ as the test statistic, defined in (5). The posterior probability has been widely used for testing and decision-making, especially for Bayesian adaptive design.[29-34] With a user-specified threshold δ_1, we reject H_0 and thus declare a safety signal at the first time point t when $P_{1,t} > \delta_1$. For example, if $\delta_1 = 0.95$, then as soon as the data evidence supports at least 95% posterior credibility of H_1, we can stop the surveillance and claim that we have detected a safety signal.

It is clear from this setup that our sequential analysis framework does *not* require a prespecified surveillance schedule as we do not precalculate a decision threshold, nor do we perform inference or make decisions based on the length or group sizes of a sequential study. Further, the posterior probability $P_{1,t}$ used for decision-making is naturally interpretable and evidence-driven, in that it quantifies how much we can trust a hypothesis based on the data accrued so far.

In Figure 3 we provide a graphical example of Bayesian sequential analysis for an exposure-outcome pair with true $RR = 2$ (ie, $\beta = \log(2)$), where analyses are performed on monthly accrued data. The shaded density curves in the top panel shows the posterior distributions $\pi_t(\beta|\mathbf{X}_t)$ inferred from monthly sequential data with the posterior median marked by an "x". As more data are accrued over time, the posterior distribution becomes more concentrated around the true effect size value, as more data evidence would reduce uncertainty. By computing the area under the density curve for which $\beta > 0$ (or $\beta \leq 0$), we can easily update the posterior probability for H_1 (or H_0) at each analysis time point. The bottom panel shows the posterior probability values updated over time. Assuming a decision threshold $\delta_1 = 0.95$, we would reject H_0 and declare a safety signal as soon as $P(H_1|\mathbf{X}_t) > 0.95$ which happens at around month 9. This suggests that by month 9, we are 95% certain that the alternative hypothesis (ie, vaccination elevates risk of the adverse event) is true given the data evidence accrued.

3.1.2 | Adaptive bias correction

At each analysis time point, we adaptively correct for residual systematic error by learning an empirical distribution for the amount of bias and then effectively "subtract" bias in a probabilistic manner to produce a posterior distribution for a "de-biased" effect size.

We do so by simultaneously analyzing a large set (typically between 50 and 100) of negative control outcomes. A negative control outcome is an outcome that is believed to have no significant association with a specific vaccine exposure.[41-43] Such outcomes are identified by the lack of evidence from reports, product labels and existing literature, and then confirmed by expert review. Intuitively, if the effect estimate of the vaccine exposure on a negative control outcome deviates

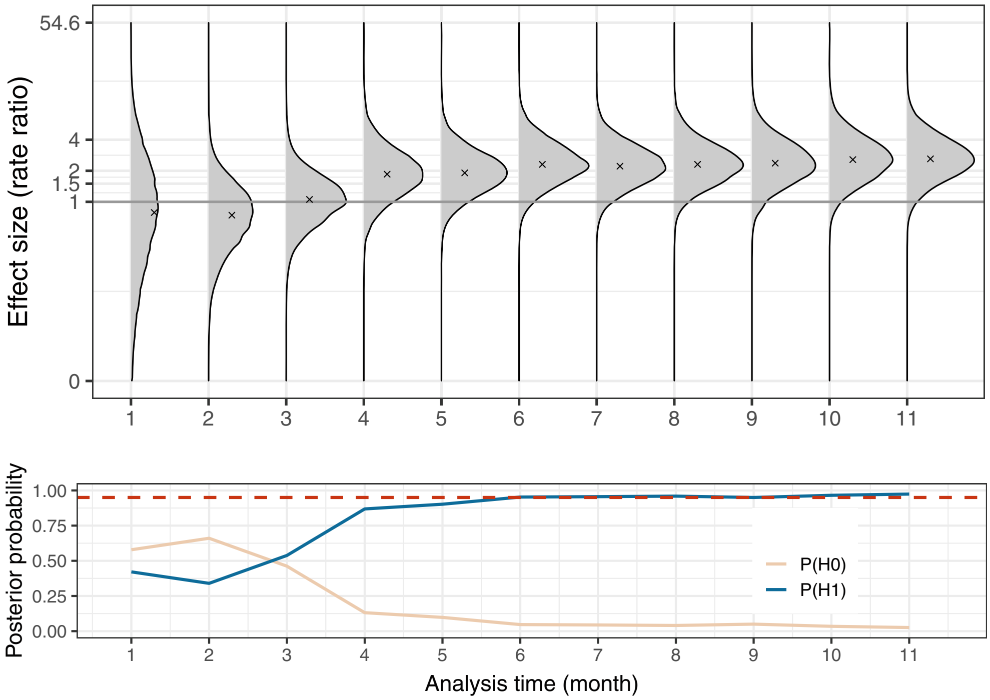


FIGURE 3 Example Bayesian sequential analysis for an exposure-outcome pair with $RR = 2$. Top: posterior distributions $\pi_t(\beta|\mathbf{X}_t)$learned with monthly sequential analyses; with more data accrued, the posterior distribution gets more concentrated around the ground truth. Bottom: Posterior probability for H_1 (blue) and H_0 (yellow) with monthly sequential analyses, directly computed from posterior distributions; with threshold $\delta_1 = 95\%$, safety signal is declared when $P_{1,t} > \delta_1$ (by month 9 in this example).

from "zero effect", then the deviation provides an empirical characterization of the residual systematic error; such effect deviations for a large set of negative control outcomes can then be taken as a "sample" to inform the underlying bias in the effect-estimate of interest.

For a specific vaccine exposure e and a set of M negative control outcomes, $\{o_i\}_{i=1}^M$, estimation of the true (unbiased) log-RR β_i related to negative control o_i could be biased by a quantity of b_i. Assuming that the bias is additive on β_i, then the biased log-RR quantity $\tilde{\beta}_i$ can be written as

$$\tilde{\beta}_i = \beta_i + b_i,$$

which, given the knowledge that $\beta_i = 0$ (since there is no association between o_i and e), indicates

$$\tilde{\beta}_i = b_i. \tag{6}$$

This suggests that profiling the estimated β for all negative control outcomes produces an empirical distribution of the bias that is due to the residual systematic error in the observational data given a specific epidemiological design.

More formally, assume that the biases b_i's associated with the negative control outcomes $\{o_i\}_{i=1}^M$ are exchangeable and follow the same distribution:

$$b_i \sim N(\overline{b}, \tau^2), \tag{7}$$

where $\overline{b}$ denotes the unknown average bias and τ^2 represents the unknown variability across different outcomes. The normal distribution could be replaced by any distribution with a density function; we illustrate our approach with the normal model here, but we have also implemented other distributions such as the t distribution with various degrees of freedom. Parameters $\overline{b}$ and τ^2 can be easily estimated by jointly fitting a normal or hierarchical normal model (or t models), given either the estimates of $\tilde{\beta}_i$'s or likelihood functions evaluated with the negative control outcomes.

Naturally, this can be done in a dynamic and adaptive manner: at time t, we can update the estimated bias distribution using all the data related to negative control outcomes accrued up to time t. That is, suppose $q_t(b|\mathbf{X}_t)$ represents the posterior predictive distribution for the bias b learned using data accrued up to time t, and we can perform bias correction for the outcome of interest by re-writing (4) with regards to the bias effect $\tilde{\beta}$:

$$\tilde{\pi}_t(\tilde{\beta}, b|\mathbf{X}_t) \propto \tilde{p}_t(\mathbf{X}_t|\tilde{\beta}) \times \tilde{\pi}_0(\tilde{\beta}) \times q_t(b|\mathbf{X}_t), \tag{8}$$

where $\tilde{\pi}_t$ denotes the joint posterior distribution for $\tilde{\beta}$ and b, $\tilde{\pi}_0$ is a prior distribution for $\tilde{\beta}$, and $\tilde{p}_t$ indicates the data likelihood function with regard to the biased effect. Using the relationship $\tilde{\beta} = \beta + b$ and thus $\beta = \tilde{\beta} - b$, inference of the true (unbiased) β is straightforward from posterior samples of $\tilde{\beta}$ and b via Markov chain Monte Carlo (MCMC).

In Figure 4 we present a graphical example of empirical bias distributions $q_t(b|\mathbf{X}_t)$ learned sequentially from monthly accrued data. The density curve by month t characterizes the learned bias distribution through negative control analysis up to month t, where each "x" marks the maximum likelihood estimate (MLE) of $\tilde{\beta}$ for each negative control outcome. Over time, more negative control estimates become available (note that more "x"s are present for later months) and the empirical bias distribution would also stabilize. In this example, a positive bias seems to persist, as the majority of the density lies above $RR = 1$ (the gray dashed line representing a null effect). We include more details of the bias correction procedure in the online Supporting Information. Furthermore, we provide a pseudocode-style sketch of the sequential analysis procedure with adaptive bias correction in Algorithm 1 and a complete, open-source implementation in the R package `EvidenceSynthesis` available at https://github.com/OHDSI/EvidenceSynthesis.

3.2 | Empirical evaluation on large-scale observational healthcare databases

The over-arching goal of our empirical evaluation is to evaluate the performance of the proposed Bayesian sequential analysis framework on real-world observational health databases, and benchmark against MaxSPRT. We use six historical vaccines with known side effects, a large set of experimental control outcomes for which the true RR values are

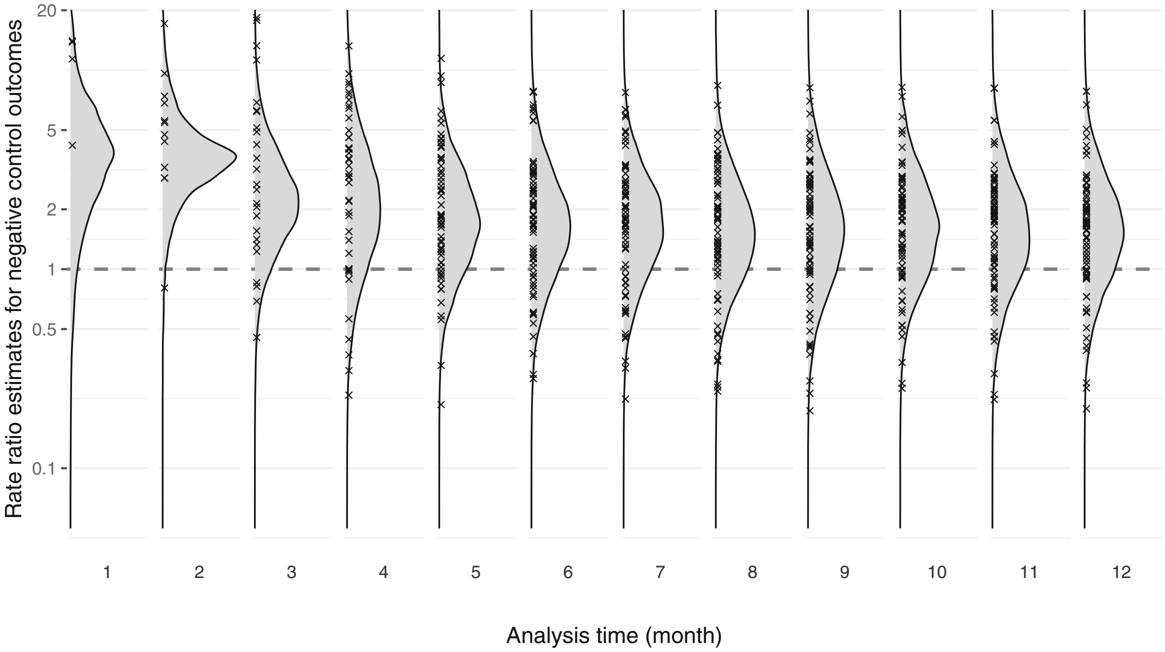


FIGURE 4 An example of learned bias distributions over time in sequential analysis. Each "x" marks the maximum likelihood estimate (MLE) of log-RR for each negative control outcome. The density curves characterize the empirical bias distribution learned from negative controls in monthly sequential analysis. The "zero-effect" value $RR = 1$ is marked by the gray dashed line; there is a positive bias since the majority of densities under these curves lie above the $RR = 1$ line.

Algorithm 1. Bayesian bias correction (BBC) for sequential analysis

```
Input: sequential data $\mathbf{X}_t$, working model likelihood $\tilde{p}$, prior $\tilde{\pi}_0$, threshold $\delta_1$ (optional)
Output: posterior distributions $\pi_t$, study stopping time $\tau_s$ (optional)
procedure BBC
    for $t = 1, 2, \dots, T_{\max}$ do                                  ▷ Sequentially analyze data
        Obtain/collect accumulated data $\mathbf{X}_t$ up to time $t$
        Split $\mathbf{X}_t$ into $\mathbf{X}_t^{(\mathrm{NC})}$ (data on negative controls) and $\mathbf{X}_t^{(\mathrm{OOI})}$ (data on outcome of interest)
        Learn empirical bias distribution $q_t(b \mid \mathbf{X}_t^{(\mathrm{NC})})$
        Obtain posterior for biased log-RR via $\tilde{\pi}_t(\tilde{\beta} \mid \mathbf{X}_t^{(\mathrm{OOI})}) \propto \tilde{p}_t(\mathbf{X}_t \mid \tilde{\beta}) \times \tilde{\pi}_0(\tilde{\beta})$
        for $s = 1, 2, \dots, S$ do                                     ▷ Learn unbiased $\beta$ via MCMC
            sample $b^{(s)} \sim q_t(b \mid \mathbf{X}_t^{(\mathrm{NC})})$
            sample $\tilde{\beta}^{(s)} \sim \tilde{\pi}_t(\tilde{\beta} \mid \mathbf{X}_t^{(\mathrm{OOI})})$
            $\beta^{(s)} \leftarrow \tilde{\beta}^{(s)} - b^{(s)}$                        ▷ Use relationship $\beta = \tilde{\beta} - b$
        Compute $P_{1,t}$, posterior probability of $H_1$ via

            $\hat{P}_{1,t} := \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}(\beta^{(s)} > 0)$          (9)

                                                  ▷ Counting fraction of MCMC samples with $\beta$>0
        if $\hat{P}_{1,t}$>$\delta_1$ then                              ▷ Stop if decision threshold crossed; optional
            Exit for loop, $\tau_s \leftarrow t$
        else if $\mathbf{X}_t$ exhausts all available data then
            Exit for loop
```

known, two commonly used analysis designs (the historical comparator and SCCS), and run the analyses on a network of observational healthcare databases mapped to Observational Health Data Science and Informatics (OHDSI,[44]) collaborative's Observational Medical Outcomes Partnership (OMOP) common data model (CDM).[45] We simulate the data accrual process by making sequential batches of data available for analysis at one-month intervals, where subject-level records become available in chronological order. We evaluate the proposed Bayesian bias correction (BBC) framework and MaxSPRT on testing-based metrics (eg, Type 1 error rates and statistical power) as well as estimation-based metrics (eg, mean squared errors).

Below we provide a brief overview of each aspect of the design of the empirical evaluation:

(a) Exposure-outcome pairs

We use six existing vaccines (or vaccine groups) including seasonal influenza, pandemic influenza (H1N1pdm), human papillomavirus (HPV), and varicella zoster virus, with data collected during specific time periods, as shown in Table 1. The starting point and end point of time periods used for the study and historical reference are recorded by "Start Date" and "End Date" under the "Study" and "Historical" columns, respectively.

Since the zoster (Shingrix) and HPV (Gardasil 9) vaccines have two doses, we split the two doses into two separate exposures and also consider a third exposure defined as receiving either dose. For instance, for the zoster (Shingrix) vaccine, we consider three exposures: zoster first dose, zoster second dose, and zoster first or second dose, where all of them share the same observational periods for the study and historical reference. Therefore, in total, we consider 10 vaccine exposures. The codes and inclusion-exclusion criteria for constructing these exposure cohorts can be found in the Appendix of the online study protocol at https://suchard-group.github.io/Better/Protocol.html#Appendix.

We use a large set of experimental control outcomes including negative control outcomes and positive control outcomes that share the systematic error structures of the negative controls. We select a single set of 93 negative control outcomes for all six vaccine groups that match the severity and prevalence of suspected vaccine adverse effects. We first generated a longer candidate list of negative controls based on similarity of prevalence and percent of diagnoses that were

TABLE 1 Vaccine exposures of interest, with start and end dates of data used for safety surveillance, and start and end dates of data used for historical reference.

	Study		Historical	
Vaccine exposure name	**Start date**	**End date**	**Start date**	**End date**
H1N1pdm	09/01/2009	05/31/2010	09/01/2008	05/31/2009
Seasonal flu (Fluvirin)	09/01/2017	05/31/2018	09/01/2016	05/31/2017
Seasonal flu (Fluzone)	09/01/2017	05/31/2018	09/01/2016	05/31/2017
Seasonal flu (All)	09/01/2017	05/31/2018	09/01/2016	05/31/2017
Zoster (Shingrix)	01/01/2018	12/31/2018	01/01/2017	12/31/2017
HPV (Gardasil 9)	01/01/2018	12/31/2018	01/01/2017	12/31/2017

recorded in an inpatient setting (as a proxy for severity) and then finalized the list with manual review by clinical experts. We use a large set of negative controls in order to represent a wide range of diseases or conditions that can cover a broad population of potential systematic errors and sources of bias. Some example negative control outcomes we have identified include chronic pancreatitis, hypothermia and leukemia.

In addition, we synthesize positive control outcomes with $RR > 1$ for which we know the true effect sizes of vaccine exposures to these control outcomes. We generate these positive controls using likelihood information of negative control outcomes and artificially injecting known effect sizes. More specifically, with respect to a particular negative control outcome, let $f_{\mathrm{nc}}(\beta)$ denote the likelihood function regarding the log-RR β. We synthesize a positive control outcome by directly synthesizing its likelihood function through *horizontally* moving $f_{\mathrm{nc}}(\beta)$ to the positive direction by a desired amount. For instance, for a positive control outcome with true effect size $\beta = \log(2)$, its synthesized likelihood function is then $f_{pc,\mathrm{RR}=2}(\beta^*) = f_{nc}(\beta - \log(2))$. In our empirical evaluation, for each negative control outcome, we synthesize three positive control outcomes, with $RR = 1.5, 2,$ and 4 (ie, $\beta = \log(1.5), \log(2)$ and $\log(4)$), respectively. We choose to use these synthesized positive control outcomes instead of real positive outcomes, as real positive outcomes are problematic for a multitude of reasons.[46] First, adverse effects of vaccines are rarely well established, and even for an established effect, the effect size (or magnitude) is never known with absolute certainty or precision. Second, for a well-known adverse effect, regulatory actions (such as restriction of use or careful monitoring) are often taken to ameliorate the risk, which will then mask such effect in real-world data.

For each exposure-outcome pair, we consider two different definitions of the time-at-risk (TAR), that is, the risk window during which a subject can experience an adverse event attributable to the vaccine exposure: (1) 1-28 days post vaccination, or (2) 1-42 days post vaccination. We estimate β separately under each TAR definition.

We also investigate Guillain-Barré syndrome (GBS) as a special outcome of interest. Previous studies have found a significant association of GBS with the zoster (Shingrix) vaccination[47] with an estimated $RR = 2.84$ within the risk window of 1-42 days after vaccination. As a brief illustration of a real-world use case of our method (in Section 5), we return to this association using the same epidemiological design and a similar data source to examine the findings made by the proposed BBC framework and by MaxSPRT.

(b) Data sources

Our evaluation uses the following US observational healthcare databases that have been widely used in previous OHDSI methodological and clinical studies (recent examples include[20,48-50]):

1. IBM MarketScan Commercial Claims and Encounters (CCAE): adjudicated health insurance claims (eg, inpatient, outpatient, and outpatient pharmacy) from large employers and health plans who provide private healthcare coverage to employees, their spouses and dependents. Population size ≈ 142 million.
2. IBM MarketScan Medicare Supplemental Database (MDCR): adjudicated health insurance claims of retirees with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service or capitated health plans. Population size ≈ 10 million.

3. IBM MarketScan Multi-State Medicaid Database (MDCD): contains adjudicated health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims. Population size ≈ 26 million.
4. Optum Clinformatics Data Mart (Clinformatics): inpatient and outpatient healthcare insurance claims for enrollees insured by commercial insurances or Medicare. Population size ≈ 85 million.
5. Optum® de-identified Electronic Health Record dataset (Optum EHR): EHRs containing clinical information, prescriptions, lab results, vital signs, body measurements, diagnoses and procedures derived from clinical notes from both inpatient and outpatient environments using natural language processing. Population size ≈ 93 million.
6. The Columbia University Irving Medical Center database (CUIMC): EHR and administrative databases including inpatient and outpatient records on demographics, visits, drugs, devices, measurements and other observations (eg, symptoms), collected from CUIMC's primary care practices in northern Manhattan and surrounding areas. Population size ≈ 6.7 million.

We provide more details on these databases in the online study protocol at https://suchard-group.github.io/Better/Protocol.html#82_Data_sources. All databases have been mapped to the OMOP CDM.[45] OHDSI's Data Quality Dashboard[51] has been used to assess and ensure data quality prior to processing any analysis results.

(c) Study designs and statistical models

We adopt two commonly used epidemiological designs to construct unexposed counterfactual person-times and compare them against exposed person-times in order to evaluate the relation between an outcome and a vaccine exposure. For each design, we also separately consider the two TAR definitions introduced in part (a).

1. Historical comparator. This is a conventional design commonly used in safety surveillance.[6,35-37,52-57] Upon computing a historical incidence rate for an adverse event on unexposed individuals during some historical time period, this design constructs the counterfactual by estimating the expected incidence rate for the present time period given the historical rate. We implement four variants with varying choices of historical time periods and covariate adjustments:

- **Unadjusted, entire year**. Using a single rate computed across the entire historical year for the entire population.
- **Age and gender adjusted, entire year**. Using a rate stratified by age (in 10 year increments) and gender, computed across the entire historical year. This allows the expected rate to be adjusted for the demographics of the vaccinated.
- **Unadjusted, time-at-risk relative to outpatient visit**. Using a single rate computed during the time-at-risk relative to a random outpatient visit in the historical year.
- **Age and gender adjusted, time-at-risk relative to outpatient visit**. Using a rate stratified by age and gender, computed during the time-at-risk relative to a random outpatient visit in the historical year.

The historical comparator design implies a Poisson count model for the adverse event counts, where the Poisson rate parameter is the incidence rate. Therefore, we use a Poisson likelihood function under this design.

2. SCCS. This is a more recently developed design that constructs counterfactual time periods for a subject who has experienced the adverse outcome (a "case") using their own trajectory.[38-40,58,59] Similarly, we implement five variants given different choices of control time periods and covariate adjustments:

- **Unadjusted SCCS excluding prevaccination window**. A simple SCCS, using all patient time when not at risk as the control time, with the exception of the 30 days prior to vaccination which is excluded from the analysis to avoid bias due to contra-indications.
- **Age & season adjusted SCCS excluding prevaccination window.** An SCCS adjusting for age and season, also excluding the 30 days prior to vaccination. Age and season will be modeled to be constant within each calendar month, and vary across months as bicubic splines.
- **Unadjusted SCCS excluding all prevaccination time**. A simple SCCS discarding all time prior to vaccination and only using post vaccination time as control time periods.
- **Self-controlled risk interval (SCRI) variant with prior control interval.** An SCRI using a control interval of 43 to 15 days prior to vaccination.

- **SCRI with post control interval**. An SCRI using a control interval of 43 to 71 days after to vaccination.

The SCCS or SCRI design implies a conditional Poisson model for the adverse event outcomes, where the incidence count during each time interval follows a (conditional) Poisson distribution with a rate parameter specific to the instantaneous risk inside the interval. Therefore, we use a conditional Poisson (or Poisson process) likelihood function under this design.

(d) Bayesian analysis choices

The Bayesian sequential analysis procedure requires the user to specify two inputs: the prior distribution π_0, and the decision threshold δ_1 for posterior probability $P_{1,t}$. We consider three choices for the prior distribution π_0 for β, all as normal distributions with mean $\mu_0 = 0$ but with different variances σ_0^2:

- **Conservative prior** with $\sigma_0^2 = 1.5$.
- **Moderately informed prior** with $\sigma_0^2 = 4$.
- **Diffuse prior** with $\sigma_0^2 = 10$.

We note that the “diffuse prior” leads to inference results that are close to maximum-likelihood estimates (MLEs) under frequentist inference.

We also consider three choices for the posterior probability threshold δ_1: 0.8, 0.9, and 0.95. However, since the choice of δ_1 has no impact on posterior estimation (but only on decision making), this threshold can also be flexibly chosen and adjusted after obtaining all inference results at the end of the surveillance period, given the retrospective nature of our evaluation study.

(e) Evaluation metrics

We benchmark the proposed BBC framework against MaxSPRT using a set of metrics to evaluate both testing and estimation performance. Note that we compute all the metrics for each framework, each design variant, each choice of risk window, and every combination of Bayesian analysis choices across all databases and vaccine exposures.

Testing-oriented metrics include:

- **Type 1 error rate (false positive rate).** Estimated as the fraction of negative controls for which a safety signal declares itself (testing statistic exceeding the threshold).
- **Type 2 error rate (false negative rate).** Estimated as the fraction of positive controls for which a safety signal is *not* declared, stratified by effect sizes of the positive control outcomes.
- **Sensitivity and specificity.** Sensitivity is equivalent to statistical power, which is 1− Type 2 error rate. Specificity is defined as 1− Type 1 error rate.
- **Time-to-detection**. The number of analyses (months) until signals are declared for a specified fraction (25% or 50%) of positive controls, stratified by effect sizes.

Given the sequential nature of the analysis, we report all testing-oriented metrics (except time to detection) measured over time.

Estimation-oriented metrics include:

- **Mean squared error (MSE)**. Mean squared error between the point estimate of β and the true β.
- **Coverage rate.** The fraction of 95% confidence or Bayesian credible intervals that cover the true β, stratified by true effect sizes of the negative or positive control outcomes.
- **Nonestimable rate.** The fraction of control outcomes for which an estimate cannot be produced.

TABLE 2 Data characteristics by exposures and databases.

Database	Exposure	Exposure	Outcome counts		Incidence rates ($\times 10^{-4}$ / p-yrs)	
	Subjects	Person-years	Median	IQR	Median	IQR
H1N1 vaccination						
CCAE	753,592	56,359.96	12.0	[6.0, 42.0]	2.13	[1.06, 7.45]
CUIMC	10,611	784.28	4.0	[2.0, 9.0]	51.00	[25.50, 114.76]
MDCD	206,865	15,447.37	4.0	[2.0, 13.5]	2.59	[1.29, 8.74]
MDCR	12,913	975.82	2.0	[1.0, 4.0]	20.50	[10.25, 40.99]
Clinformatics	457,565	34,373.75	13.0	[5.0, 35.0]	3.78	[1.45, 10.18]
Optum EHR	156,467	11,962.63	5.0	[2.0, 11.0]	4.18	[1.67, 9.20]
Seasonal flu vaccination (Fluvirin)						
CCAE	119,186	9,022.36	4.0	[2.0, 11.0]	4.43	[2.22, 12.19]
CUIMC	230	17.39	1.0	[1.0, 1.0]	575.20	[575.20, 575.20]
MDCD	15,282	1,160.04	2.0	[1.0, 6.2]	17.24	[8.62, 53.88]
MDCR	822	61.75	1.0	[1.0, 1.0]	161.93	[161.93, 161.93]
Clinformatics	189,184	14,372.50	7.0	[3.0, 14.8]	4.87	[2.09, 10.26]
Optum EHR	14,706	1,120.67	2.0	[1.0, 4.0]	17.85	[8.92, 35.69]
Seasonal flu vaccination (Fluzone)						
CCAE	957	69.60	1.0	[1.0, 1.0]	143.67	[143.67, 143.67]
CUIMC	3,397	260.37	1.0	[1.0, 3.5]	38.41	[38.41, 134.42]
MDCD	3,357	256.48	2.0	[1.0, 2.0]	77.98	[38.99, 77.98]
MDCR	34,414	2,615.74	2.0	[1.0, 5.5]	7.65	[3.82, 21.03]
Clinformatics	798,816	61,118.75	40.5	[12.0, 96.8]	6.63	[1.96, 15.83]
Optum EHR	337,218	23,690.50	12.5	[5.0, 28.0]	5.28	[2.11, 11.82]
Seasonal flu vaccination (All)						
CCAE	3,516,811	266,796.18	78.5	[27.5, 214.8]	2.94	[1.03, 8.05]
CUIMC	117,339	8,939.38	12.0	[5.0, 39.0]	13.42	[5.59, 43.63]
MDCD	1,237,934	94,034.65	40.0	[19.0, 136.0]	4.25	[2.02, 14.46]
MDCR	264,636	20,078.82	18.0	[7.8, 43.5]	8.96	[3.86, 21.66]
Clinformatics	3,399,471	258,963.54	166.0	[52.8, 330.8]	6.41	[2.04, 12.77]
Optum EHR	2,536,334	190,273.53	100.0	[42.0, 251.0]	5.26	[2.21, 13.19]
First HPV vaccination (Gardasil 9)						
CCAE	376,341	28,202.88	6.0	[2.0, 21.0]	2.13	[0.71, 7.45]
CUIMC	7,670	574.87	1.0	[1.0, 2.5]	17.40	[17.40, 43.49]
MDCD	236,683	17,767.07	4.0	[1.0, 21.0]	2.25	[0.56, 11.82]
Clinformatics	173,228	12,938.73	6.0	[1.2, 9.0]	4.64	[0.97, 6.96]
Optum EHR	233,985	17,301.64	5.0	[2.0, 15.5]	2.89	[1.16, 8.96]
MDCR	0	0				
Second HPV vaccination (Gardasil 9)						
CCAE	49,283	3,486.95	2.0	[1.0, 4.0]	5.74	[2.87, 11.47]
CUIMC	1,172	84.26	1.0	[1.0, 1.5]	118.68	[118.68, 178.03]
MDCD	15,065	1,066.11	2.0	[1.2, 4.0]	18.76	[11.72, 37.52]
Clinformatics	21,377	1,508.51	2.0	[1.0, 4.0]	13.26	[6.63, 26.52]
Optum EHR	28,336	2,005.88	2.0	[1.0, 3.0]	9.97	[4.99, 14.96]
MDCR	0	0				

TABLE 2 (Continued)

	Exposure	Exposure	Outcome counts		Incidence rates ($\times 10^{-4}$/ p-yrs)	
Database	Subjects	Person-years	Median	IQR	Median	IQR
First or second HPV vaccination (Gardasil 9)						
CCAE	378,052	31,689.85	6.0	[2.0, 22.5]	1.89	[0.63, 7.10]
CUIMC	7,726	659.14	1.0	[1.0, 2.5]	15.17	[15.17, 37.93]
MDCD	237,455	18,833.17	4.0	[1.0, 21.8]	2.12	[0.53, 11.55]
Clinformatics	174,692	14,447.25	6.0	[1.8, 10.2]	4.15	[1.21, 7.09]
Optum EHR	234,518	19,307.52	5.5	[2.0, 16.8]	2.85	[1.04, 8.68]
MDCR	0	0				
First zoster vaccination (Shingrix)						
CCAE	148,190	11,004.20	4.0	[2.0, 13.0]	3.63	[1.82, 11.81]
CUIMC	11,182	835.11	2.0	[1.0, 4.5]	23.95	[11.97, 53.88]
MDCD	11,407	851.97	2.0	[1.0, 6.0]	23.48	[11.74, 70.43]
MDCR	52,789	3,952.34	4.0	[2.0, 9.8]	10.12	[5.06, 24.67]
Clinformatics	229,463	17,113.06	10.0	[5.0, 27.0]	5.84	[2.92, 15.78]
Optum EHR	219,665	16,251.16	8.0	[3.0, 27.0]	4.92	[1.85, 16.61]
Second zoster vaccination (Shingrix)						
CCAE	72,063	5,117.29	3.0	[1.5, 7.5]	5.86	[2.93, 14.66]
CUIMC	4,229	307.32	1.5	[1.0, 2.8]	48.81	[32.54, 89.48]
MDCD	5,379	388.03	1.0	[1.0, 3.0]	25.77	[25.77, 77.31]
MDCR	30,218	2,161.22	3.0	[1.0, 7.0]	13.88	[4.63, 32.39]
Clinformatics	119,556	8,506.70	6.0	[2.0, 13.0]	7.05	[2.35, 15.28]
Optum EHR	63,464	4,585.85	4.0	[2.0, 9.0]	8.72	[4.36, 19.63]
First or second zoster vaccination (Shingrix)						
CCAE	149,219	16,121.49	6.0	[3.0, 13.8]	3.72	[1.86, 8.53]
CUIMC	11,211	1,142.45	2.0	[1.0, 6.0]	17.51	[8.75, 52.52]
MDCD	11,556	1,239.99	2.0	[2.0, 7.0]	16.13	[16.13, 56.45]
MDCR	53,384	6,113.56	5.0	[2.0, 15.0]	8.18	[3.27, 24.54]
Clinformatics	232,669	25,619.78	13.0	[5.0, 36.0]	5.07	[1.95, 14.05]
Optum EHR	220,106	20,837.00	10.0	[4.0, 32.5]	4.80	[1.92, 15.60]

Note: Column "Exposure Subjects" shows the total number of unique people with vaccination exposure considered in the analysis. Column "Exposure Days" shows the cumulative at-risk days for all exposure subjects. Column "Outcome counts" shows the median event counts across 93 negative control outcomes during all exposure time periods; numbers in the parentheses are the 25th and 75th percentiles (ie, the interquartile range, IQR). Column "Incidence rates" shows the incidence rate per person-year (incident count divided by exposure person-years) across all 93 negative control outcomes, similarly with the median and IQR. Summary is presented for design choices with 1-28 days after vaccine exposure considered as the "time-at-risk".

(f) Data characteristics overview

We further present some key data characteristics in Table 2. Here we summarize, for each vaccine exposure and each database, the total number of subjects exposed, the total accumulated exposure time (in person-years), and some summary statistics (median and inter-quartile range) of the incident count and incident rate across all negative control outcomes, with the "time-at-risk" taken as 1-28 days post vaccination. More data characteristics information is provided in the online Supporting Information.

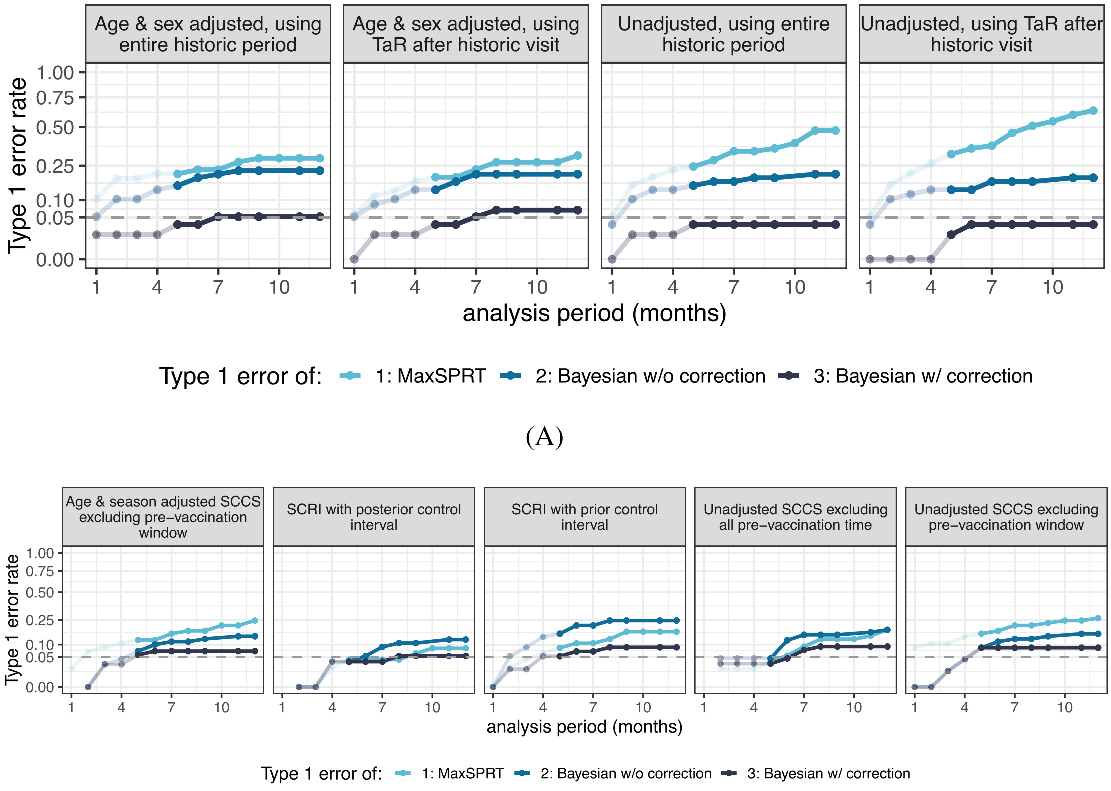
4 | EVALUATION RESULTS

In this section we discuss results from the empirical evaluation comparing the performance of our proposed sequential analysis framework with Bayesian bias correction (BBC) and that of MaxSPRT. All MaxSPRT analyses are implemented with significance level $\alpha = 0.05$, and with a prespecified surveillance schedule that *exactly matches* the actual data accrual process—this is technically impossible in practice, and such a choice is favorable to MaxSPRT and should produce the best possible performance of using MaxSPRT for safety surveillance on real-world data.

Since there are more than 7 000 000 sets of analysis results from this large-scale empirical evaluation, we present a collection of representative results in terms of the test-oriented metrics, and provide summary statistics in terms of the estimation-oriented metrics. If not otherwise specified, in this section we focus on Bayesian analyses using the "moderately informed prior" with prior variance $\sigma_0^2 = 4$. We choose not to report on studies with insufficient evidence where the maximum incidence count across all negative control outcomes is lower than 2, after examining the data characteristics (Table 2) but before inspecting the results. For completeness, we make all results publicly available through an R ShinyApp at https://data.ohdsi.org/BetterExplorer.

4.1 | The Bayesian framework controls Type 1 error better

In Figure 5, as a typical example, we plot the empirical Type 1 error rates over analysis time-points (in months) using the Bayesian framework and MaxSPRT and for the HPV (Gardasil 9) vaccine exposures within the CCAE database with 1-28



(A)

(B)

FIGURE 5 The Bayesian framework offers better Type 1 error control than MaxSPRT. Empirical Type 1 error rates over analysis period (in months) using proposed Bayesian methods and MaxSPRT. Results are shown for HPV vaccine exposure (after first or second dose of Gardasil 9) within the CCAE database, with 1-28 days post vaccination considered as the time-at-risk. (A) Type 1 error rates for historical comparator designs. Each panel shows results for a different design variant. (B) Type 1 error rates for SCCS designs. Each panel shows results for a different design variant.

days post vaccination as the time-at-risk. Within the Bayesian framework, we have also implemented a baseline with the bias correction component removed (dark blue curves); for both the baseline and the full approach with bias correction, we adopt a decision threshold $\delta_1 = 0.95$. We remind the reader that *a lower Type 1 error rate is desired* as one generally wants to avoid producing excessive false positive signals. Further, we have intentionally made results from the first 4 months more transparent to highlight later analysis periods, since in the early phases data remain insufficient to deliver stable estimates.

It is clear that the proposed sequential analysis procedure with BBC (dark curves) offers substantially better control of Type 1 error rates compared to MaxSPRT (light blue curves). This is shared between historical comparator designs (subplot (a)) and SCCS designs (subplot (b)), but the difference is more prominent with historical comparator designs—note that for the two unadjusted designs the Type 1 error rates that MaxSPRT achieves even exceed 50%, 10 times higher than the prespecified $\alpha = 5\%$. Even though the Bayesian framework does not rely on a prespecified significance level, with adaptive bias correction, it actually provides a near 5% actual Type 1 error rate across design choices; in many cases, Bayesian sequential analysis even without bias correction still out-performs MaxSPRT in this metric, although the benefit is not as large as those offered by the full Bayesian approach.

4.2 | The Bayesian framework provides higher power

We next examine the statistical power in identifying true safety signals. Since our Bayesian approach and MaxSPRT return notably different Type 1 error rates and we wish to make more informative comparisons, we first numerically select a decision threshold under each epidemiological design such that the end-of-analysis Type 1 error rate that the Bayesian approach returns approximately equals to the end-of-analysis Type 1 error rate from MaxSPRT. In this way, we are comparing their statistical power when the same amount of Type 1 error is allowed between the Bayesian approach and MaxSPRT.

In Figure 6, we present results for one typical historical comparator design and one SCCS design in Figure 6, with the same vaccine exposure and database as in Figure 5. Despite the intrinsic trade-off between Type 1 error rate and statistical power, with the same allowance on Type 1 error, our Bayesian framework offers greater power over MaxSPRT in most scenarios, while delivering comparable power in other scenarios. We note that this gain is more obvious with smaller effect sizes (eg, $RR = 1.5$ or $RR = 2$) that, likewise, are more representative for adverse events associated with vaccination. This means that our proposed Bayesian approach is able to capture more true positive safety signals, while maintaining the same level of false positive decisions as MaxSPRT.

We further validate this point by inspecting the timeliness of identifying true positive signals. In Figure 7, we compare the time-to-detection, a method takes to declare safety signals for at least 50% of all positive control outcomes, between our proposed BBC procedure and MaxSPRT, while allowing for approximately the same amount of Type 1 error. Figure 5 presents this comparison for all epidemiological designs, again, for HPV vaccine exposure in the CCAE database with 1-28 days post vaccination as the time-at-risk. Since the goal of safety surveillance is rapid detection of safety concerns, a *shorter time-to-detection is desired*. Across almost all designs, the proposed Bayesian approach takes shorter time to detect at least 50% of the true positive controls, particularly for smaller effect sizes. In few cases, the Bayesian approach does take slightly, but not substantially, longer time.

4.3 | Bayesian bias correction yields more accurate estimation

As the proposed Bayesian framework directly targets and corrects for estimation bias, it is able to produce more accurate and reliable estimates of log-*RR*, β. In Table 3, we present the mean-squared errors (MSEs) in estimating β, using our proposed Bayesian BBC approach and maximum-likelihood estimation (MLE) under MaxSPRT. For each combination of database, exposure, epidemiological design, time-at-risk choice, and true effect size β of control outcomes, we take the average of the squared estimation errors across outcomes to produce one estimate of MSE. Table 3 summarizes the distribution of all those MSEs by examining their overall average, median, as well as the 10th, 25th, 75th, and 90th percentiles. *Lower MSEs are desired* for estimation. It is clear that across the spectrum of all analyses, the BBC has much lower estimation error compared to MLE under MaxSPRT. Notably, on average, BBC yields point estimates with a nearly 80% MSE reduction compared to MLE under MaxSPRT.

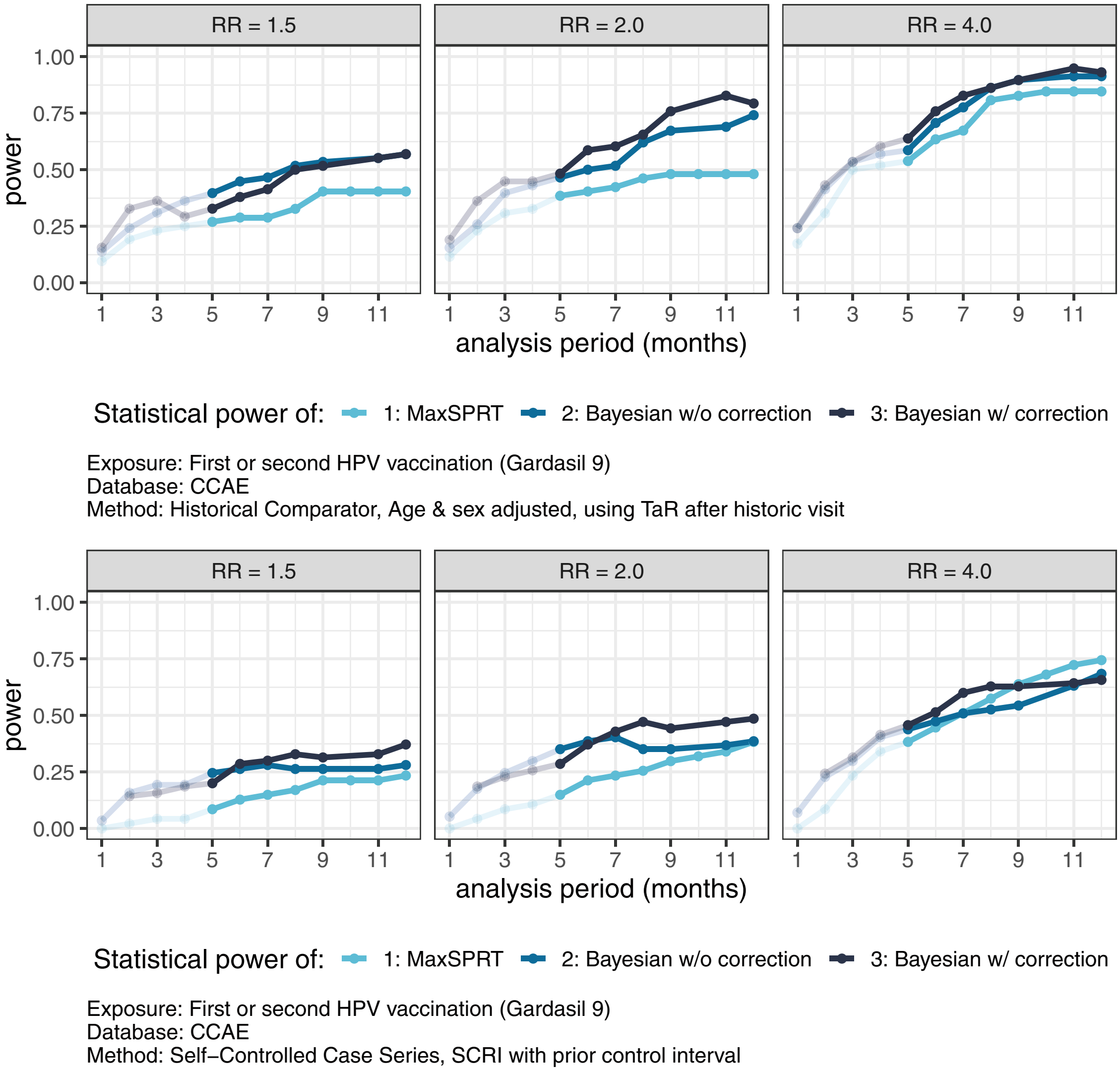
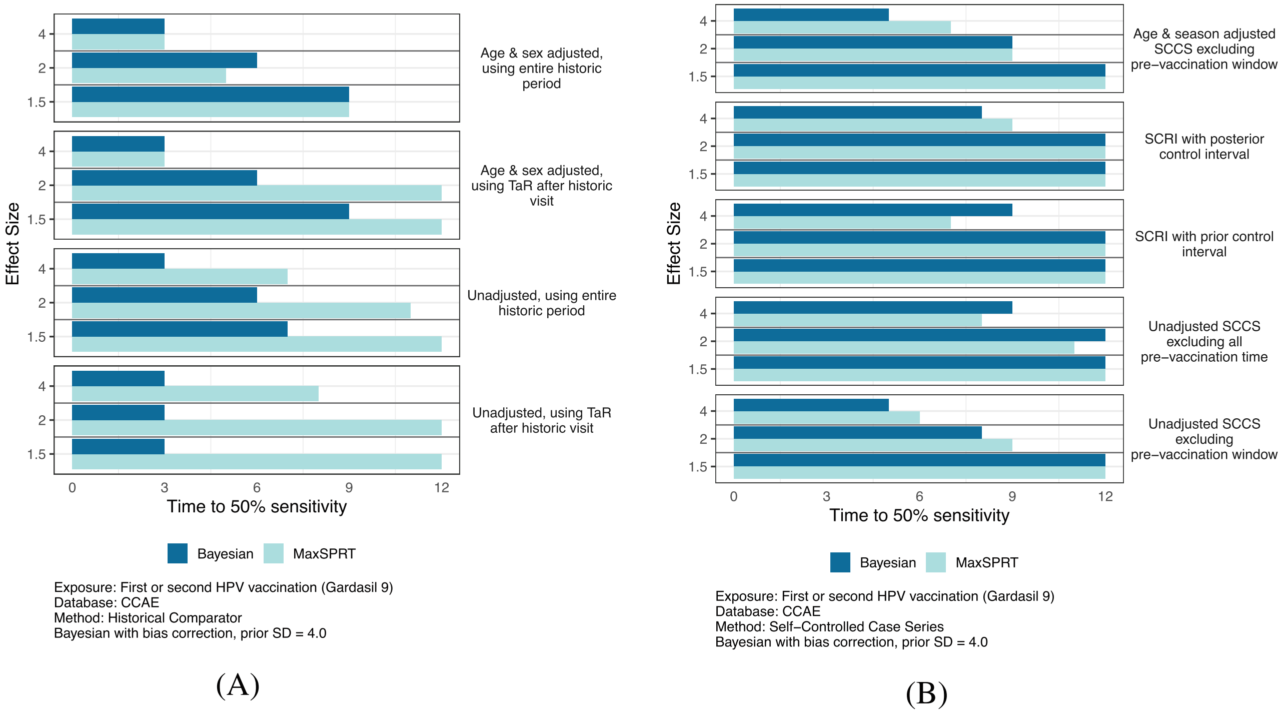


FIGURE 6 Bayesian framework offers higher statistical power. Statistical power (= 1− Type 2 error rate) over analysis periods (in months) using the proposed Bayesian methods and MaxSPRT, stratified by effect sizes of positive control outcomes, with $RR = 1.5, 2$, and 4. Results are shown for HPV vaccine exposure using one representative historical comparator design (Top) and one representative SCCS design (Bottom) within the CCAE database. The Bayesian method with bias correction almost consistently produces higher statistical power over MaxSPRT, when the same empirical Type 1 error level is allowed across methods.

Taking a similar approach to examine coverage rates of 95% credible/confidence intervals, we present a summary of coverage study in Table 4 in terms of the average, median, and 25th and 75th percentiles (inter-quartile range) of coverage rates calculated across all analyses. BBC returns 95% credible intervals that offer close-to-95% coverage of the true β values very consistently. At the meantime, 95% confidence intervals using MLE under MaxSPRT fail to provide nominal coverage for more than 75% of all confidence intervals (note that all the 75th percentiles are even below 0.95).

5 | REAL-DATA ILLUSTRATION: GUILLAIN-BARRÉ SYNDROME RISK POST ZOSTER VACCINATION

To illustrate a use case of our proposed surveillance procedure, we present results from a brief example study investigating the association between occurrences of Guillain-Barré syndrome (GBS) and exposure to either of the two doses of the zoster (Shingrix) vaccine on a commonly used observational data source. Previous work[47] suggests that the zoster (Shingrix) vaccine induces an increased risk of GBS post vaccination, with an estimate of $RR = 2.84$



Exposure: First or second HPV vaccination (Gardasil 9)
Database: CCAE
Method: Historical Comparator
Bayesian with bias correction, prior SD = 4.0

Exposure: First or second HPV vaccination (Gardasil 9)
Database: CCAE
Method: Self−Controlled Case Series
Bayesian with bias correction, prior SD = 4.0

FIGURE 7 Timeliness: Bayesian framework detects true positive signals faster. Time-to-detection it takes to reach 50% sensitivity by the proposed BBC framework and MaxSPRT, with the same level of empirical Type 1 error rates. Results are shown for HPV vaccine exposures within the CCAE database, with 1-28 days post vaccination as time-at-risk. (A) Historical comparator designs. (B) Self-controlled case series designs.

TABLE 3 Bayesian bias correction (BBC) provides more accurate estimation.

	MSE (summary quantiles)					
	Average	**10%**	**25%**	**Median**	**75%**	**90%**
BBC	0.709	0.170	0.245	0.408	0.911	1.596
MLE	3.394	0.214	0.340	0.597	3.218	8.107

Note: Summary of mean squared errors (MSEs) for estimating β, comparing BBC and MLE under MaxSPRT. MSEs are calculated for different effect sizes of control outcomes regarding each exposure within each epidemiological design on each database, for 3352 total analyses where MLE under MaxSPRT is able to produce estimates.

(95% CI: 1.53, 5.27) using observational data from Medicare data sources. We wish to examine this question within the MDCR database (a subset of Medicare) by applying the same epidemiological design used in Reference 47, but comparing the performance of different statistical approaches, namely the proposed Bayesian framework and MaxSPRT.

Table 5 presents the final estimates obtained after 12 analyses based on monthly accrued data. Neither the Bayesian approach nor MaxSPRT has detected a positive signal that suggests zoster vaccination elevates the risk of GBS throughout the sequential analyses. Our Bayesian framework does return an estimate of $RR = 2.82$ that is very close to the estimate produced in previous studies,[47] while MLE under MaxSPRT seems to over-estimate the effect. However, the credible/confidence intervals are very wide, indicating inconclusive results. This is because the incidence counts of GBS within MDCR stand very low: only 6 total cases of GBS (defined as subjects experiencing any instance of GBS during the observation period) were present, compared to the 44 total cases in the Medicare data sources used previously.[47] Although our estimates remain inconclusive, they do highlight the potential that Bayesian safety surveillance with BBC could replicate and validate previous findings if more data were available. A simple power analysis suggests that we would need at least 20 total GBS cases in our data source to detect a statistically significant signal.

TABLE 4 BBC produces credible intervals with substantially higher coverage rates than MLE under MaxSPRT-produced confidence intervals.

Coverage rates of 95% intervals (summary)				
	Average	**25%**	**Median**	**75%**
Effect size $\beta = \log(1)$				
BBC	0.953	0.941	0.962	0.980
MLE	0.682	0.480	0.811	0.918
Effect size $\beta = \log(1.5)$				
BBC	0.948	0.938	0.959	0.978
MLE	0.677	0.467	0.801	0.934
Effect size $\beta = \log(1.5)$				
BBC	0.943	0.929	0.957	0.976
MLE	0.677	0.467	0.801	0.934
Effect size $\beta = \log(4)$				
BBC	0.925	0.909	0.944	0.967
MLE	0.677	0.467	0.801	0.934

Note: Summary of coverage rates of 95% credible/confidence intervals produced by BBC and MLE under MaxSPRT in estimating β, stratified by the true effect sizes. Across all effect sizes, the proposed Bayesian method produces 95% credible intervals that with close to 95% empirical coverage rates, while 95% confidence intervals produced by MLE under MaxSPRT tend to under-cover. Empirical coverage rates are calculated for different effect sizes of control outcomes regarding each exposure within each design on each database, for 3352 total analyses where MLE under MaxSPRT is able to produce estimates.

TABLE 5 *RR* estimates for occurrence of Guillain-Barré syndrome (GBS) with versus without zoster (Shingrix) vaccinations.

	Estimate		**Case counts**	
Method	***RR***	**95% CI**	**Risk**	**Control**
BBC (proposed)	2.82	(0.430, 19.0)	2	4
MLE (MaxSPRT)	4.85	(0.562, 41.6)	2	4
Goud et al. (2021)	2.84	(1.53, 5.27)	24	20

Note: A zoster (Shingrix) vaccine exposure is considered as taking either of the two doses. A SCCS design with a postvaccination control window and covariate adjustment is applied on the MDCR (Medicare) database. We compare our Bayesian framework with BBC and MLE under MaxSPRT estimates against the estimates reported in previous studies.[47] Columns "Risk" and "Control" under "Case Counts" record the numbers of all GBS cases during the risk interval (1-42 days post vaccination) and during the control interval (43-183 days post vaccination).

6 | DISCUSSION

This article proposes a Bayesian sequential analysis framework for vaccine safety surveillance with adaptive bias correction. Our framework delivers a unified statistical solution to simultaneously provide a flexible surveillance schedule and correct for bias induced by residual systematic error in observational data. Our approach relies on accrued data only without the need to prespecify a surveillance schedule, as it summarizes data evidence through posterior distributions of effect sizes and uses posterior probabilities of hypotheses for sequential testing. Therefore, unlike MaxSPRT, our Bayesian framework is more flexible and adaptive to practical data settings as it does not depend on a priori predictions on data accrual behaviors. Furthermore, we address the challenge of residual systematic error and confounding with a joint statistical model that adaptively learns and corrects for bias by simultaneously analyzing a large set of negative control outcomes through Bayesian hierarchical modeling. This data-driven approach enables us to substantially reduce estimation bias and remedy testing error inflation that MaxSPRT suffers from. Notably, our procedure differs from conventional two-stage surveillance approaches[60] and enables more reliable safety signal generation that no longer requires separate signal validation and can fully utilize available data for improved statistical power. Through a comprehensive empirical evaluation on six large-scale, real-world healthcare databases covering more than 360 million unique patients, we demonstrate that the proposed framework offers better control of Type 1 error, high statistical power, fast detection of safety signals, and more accurate and reliable estimation.

There are, admittedly, several limitations in this work. First, our adaptive bias correction approach assumes exchangeable biases in the negative controls, though exchangeability is conditional on data source, epidemiological design and analysis time. Second, we do not explicitly model time-varying risks of adverse events and do not investigate time-varying confounding, although our sequential adaptive bias correction procedure implicitly remedies this issue in part. Third, we have only adopted two commonly used epidemiological designs, the historical comparator and self-controlled case series, which may not be the most suitable design for vaccine safety surveillance situations with complex roll-out schedules (eg, COVID-19 vaccines). Finally, as misclassification of study variables is unavoidable in secondary use of healthcare data, it is possible to misclassify exposures, covariates, and outcomes; we do not expect differential misclassification, so bias will most likely be towards the null.

Nonetheless, our empirical results lead to several interesting future directions. First of all, with auxiliary information on correlation structures between negative control outcomes (eg, via clinical expert review), we can extend our framework with a hierarchical mixture modeling approach to allow for nonexchangeable bias distributions. Secondly, since the Bayesian framework no longer requires a prefixed surveillance schedule, we can motivate theoretical investigation into long-term error control with infinite time horizons in sequential testing for composite hypotheses. Further, motivated by the inconclusive case study, we can develop meta-analysis approaches for synthesizing evidence across multiple data sources to increase statistical power for safety signal detection, particularly on rare safety outcomes. Finally, we can perform a follow-up empirical evaluation using other epidemiological designs and with more recently approved vaccines with complex roll-out schedules.

ACKNOWLEDGEMENTS

FB, AN, KK, GH, and MAS receive a contract from the US Food & Drug Administration to support this work. Further, FB and MAS receive grants and contracts from the US National Institutes of Health (NIH), the US Department of Veterans Affairs and Janssen Research & Development outside the scope of this work. AN is partially supported by NIH R01 AG068002 and GH is partially supported by NIH R01 LM006910. PBR and MJS are employees of Janssen Research & Development and shareholders in Johnson & Johnson.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Fan Bu https://orcid.org/0000-0003-3697-1477
Martijn J. Schuemie https://orcid.org/0000-0002-0817-5361

REFERENCES

1. Anderson Roy M, May RM. Directly transmitted infections diseases: control by vaccination. *Science*. 1982;215:1053-1060.
2. Francis Donald P, Hadler Stephen C, Thompson Sumner E, et al. The prevention of hepatitis B with vaccine: report of the Centers for Disease Control multi-center efficacy trial among homosexual men. *Ann Intern Med*. 1982;97:362-366.
3. Centers for Disease Control and Prevention. National Immunization Program (Centers for Disease Control and Prevention), Education and Information and Partnership Branch National Immunization Program (Centers for Disease Control and Prevention). *Epidemiology and Prevention of Vaccine-Preventable Diseases*. Atlanta, GA: Department of Health & Human Services, Public Health Service, Centers for Disease Control and Prevention; 2005.
4. Nowak Glen J, Kristine S, Kelli B, Smith Teresa M, Michelle B. Promoting influenza vaccination: insights from a qualitative meta-analysis of 14 years of influenza-related communications research by US Centers for Disease Control and Prevention (CDC). *Vaccine*. 2015;33:2741-2756.
5. Thomas V, Frank DS, Chen Robert T, Elizabeth M. Vaccine safety surveillance using large linked databases: opportunities, hazards and proposed guidelines. *Expert Rev Vaccines*. 2003;2:21-29.
6. Lieu Tracy A, Martin K, Davis Robert L, et al. Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care*. 2007;45:S89-S95.
7. Baker Meghan A, Michael N, Cole David V, Lee Grace M, Lieu TA. Post-licensure rapid immunization safety monitoring program (PRISM) data characterization. *Vaccine*. 2013;31:K98-K112.

8. Andreia L, Andrews Nick J, Thomas SL. Near real-time vaccine safety surveillance using electronic health records—a systematic review of the application of statistical methods. *Pharmacoepidemiol Drug Saf*. 2016;25:225-237.
9. Moro Pedro L, Rongxia L, Penina H, Eric W, Maria C. Surveillance systems and methods for monitoring the post-marketing safety of influenza vaccines at the Centers for Disease Control and Prevention. *Expert Opin Drug Saf*. 2016;15:1175-1183.
10. Lee Grace M, Romero José R, Bell BP. Postapproval vaccine safety surveillance for COVID-19 vaccines in the US. *JAMA*. 2020;324:1937-1938.
11. EMA European Medicine Agency. Pharmacovigilance Plan of the EU Regulatory Network for COVID-19 Vaccines (EMA/333964/2020). 2020. https://www.ema.europa.eu/en/documents/other/pharmacovigilance-plan-eu-regulatory-network-covid-19-vaccines_en.pdf
12. Centers for Disease Control and Prevention VAERS. Vaccine Adverse Event Reporting System (VAERS). 2021. https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vaers/index.html
13. Centers for Disease Control and Prevention VSD. Vaccine Safety Datalink (VSD). 2021. https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vsd/index.html
14. Centers for Disease Control and Prevention CISA. Clinical Immunization Safety Assessment (CISA) Project. 2021. https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/cisa/index.%html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fvaccinesafety%2Factivities2FCISA.html
15. World Health Organization (WHO). Establishing Surveillance Systems in Countries Using COVID-19 Vaccines. 2021. https://cdn.who.int/media/docs/default-source/covid-19-vaccines-safety-surveillance-manual/training-slides_covid-19_vs_surveillance_systemsafa53e9c-0bde-4765-90fa-2dedf3b6da72.pdf?sfvrsn=6ccff509_5&Status=Master
16. Rodrigues LC, Smith PG. Use of the case-control approach in vaccine evaluation: efficacy and adverse effects. *Epidemiol Rev*. 1999;21:56-72.
17. Glanz Jason M, McClure David L, Stanley X, et al. Four different study designs to evaluate vaccine safety were equally validated with contrasting limitations. *J Clin Epidemiol*. 2006;59:808-818.
18. Newcomer Sophia R, Martin K, Stan X, et al. Bias from outcome misclassification in immunization schedule safety research. *Pharmacoepidemiol Drug Saf*. 2018;27:221-228.
19. Martin K, Davis RL, Margarette K, Edwin L, Tracy L, Richard P. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. *Seq Anal*. 2011;30:58-78.
20. Schuemie Martijn J, Faaizah A, Nicole P, et al. Vaccine Safety Surveillance Using Routinely Collected Healthcare Data—An Empirical Evaluation of Epidemiological Designs. *Front Pharmacol*. 2022;13:2532.
21. Schuemie Martijn J, Fan B, Akihiko N, Suchard Marc A. Adjusting for both sequential testing and systematic error in safety surveillance using observational data: Empirical calibration and MaxSPRT. *Stat Med*. 2023;42:619-631.
22. Barnard GA. Sequential tests in industrial statistics. *Suppl J R Stat Soc*. 1946;8:1-21.
23. Wetherill GB. Bayesian sequential analysis. *Biometrika*. 1961;48:281-292.
24. Berger James O, Brown Lawrence D, Wolpert RL. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann Stat*. 1994;22:1787-1807.
25. Berger James O, Wolpert RL. The likelihood principle. *Institute of Mathematical Statistics Lecture Notes–Monograph*. Beachwood, OH: IMS; 1988.
26. Berger James O, Ben B, Yinping W. Unified frequentist and Bayesian testing of a precise hypothesis. *Stat Sci*. 1997;12:133-160.
27. Berger James O, Benzion B, Yinping W. Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*. 1999;86:79-92.
28. Jha Sumit K, Clarke Edmund M, Langmead Christopher J, Axel L, André P, Paolo Z. A bayesian approach to model checking biological systems. *International Conference on Computational Methods in Systems Biology*. Hanover, PA: Springer; 2009:218-234.
29. Jerome C. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J Am Stat Assoc*. 1966;61:577-594.
30. Thall Peter F, Simon Richard M, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med*. 1995;14:357-379.
31. Smith Michael K, Ieuan J, Morris Mark F, Grieve Andrew P, Keith T. Implementation of a Bayesian adaptive design in a proof of concept study. *Pharm Stat J Appl Stat Pharm Ind*. 2006;5:39-50.
32. Xian Z, Suyu L, Kim Edward S, Herbst Roy S, Jack LJ. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clin Trials*. 2008;5:181-193.
33. Berry Scott M, Carlin Bradley P, Jack LJ, Peter M. *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC press; 2010.
34. Rongxia L, Brock S, Charles R. A Bayesian approach to sequential analysis in post-licensure vaccine safety surveillance. *Pharm Stat*. 2020;19:291-302.
35. Belongia Edward A, Irving Stephanie A, Shui Irene M, et al. Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. *Pediatr Infect Dis J*. 2010;29:1-5.
36. Xintong L, Anna O, Rupa M, et al. Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. *BMJ*. 2021;373.
37. Klein Nicola P, Ned L, Kristin G, et al. Surveillance for adverse events after COVID-19 mRNA vaccination. *JAMA*. 2021;326:1390-1399.
38. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med*. 2006;25:1768-1797.
39. Paddy F, Rush M, Miller E, et al. A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. *Lancet*. 1995;345:567-569.
40. Farrington CP. Control without separate controls: evaluation of vaccine safety using case-only methods. *Vaccine*. 2004;22:2064-2070.

41. Tchetgen E. The control outcome calibration approach for causal inference with unobserved confounding. *Am J Epidemiol*. 2014;179:633-640.
42. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med*. 2014;33:209-218.
43. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. *Stat Med*. 2016;35:3883-3888.
44. George H, Duke Jon D, Shah Nigam H, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *MEDINFO 2015: eHealth-enabled Health*. Amsterdam, the Netherlands: IOS Press; 2015:574-578.
45. Overhage J, Marc RPB, Reich Christian G, Hartzema Abraham G, Stang Paul E. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012;19:54-60.
46. Schuemie Martijn J, Ryan Patrick B, George H, David M, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Transact A Math Phys Eng Sci*. 2018;376:20170356.
47. Ravi G, Bradley L, Jonathan D, et al. Risk of Guillain-Barré Syndrome Following Recombinant Zoster Vaccine in Medicare Beneficiaries. *JAMA Intern Med*. 2021;181:1623-1630.
48. Anna O, George H. COVID-19 vaccination effectiveness rates by week and sources of bias: a retrospective cohort study. *BMJ Open*. 2022;12:e061126.
49. Suchard Marc A, Schuemie Martijn J, Krumholz Harlan M, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet*. 2019;394:1816-1826.
50. George H, Ryan Patrick B, Duke Jon D, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences*. Washington, D.C: National Academy of Sciences, Vol 113; 2016:7329-7336.
51. Clair B, Defalco Frank J, Ryan Patrick B, Rijnbeek Peter R. Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc*. 2021;28:2251-2257.
52. Steven B, Juhani E, Claire-Anne S, et al. Importance of background rates of disease in assessment of vaccine safety during mass immunisation with pandemic H1N1 influenza vaccines. *Lancet*. 2009;374:2115-2122.
53. Katherine YW, Nordin James D, Martin K, et al. An assessment of the safety of adolescent and adult tetanus–diphtheria–acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the Vaccine Safety Datalink. *Vaccine*. 2009;27:4257-4262.
54. Katherine YW, Martin K, Fireman Bruce H, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. *Pediatrics*. 2011;127:S54-S64.
55. Buttery JP, Danchin MH, Lee KJ, et al. Intussusception following rotavirus vaccine administration: post-marketing surveillance in the National Immunization Program in Australia. *Vaccine*. 2011;29:3061-3066.
56. Leonoor W, Coralie L, Corinne V, et al. The incidence of narcolepsy in Europe: before, during, and after the influenza A (H1N1) pdm09 pandemic and vaccination campaigns. *Vaccine*. 2013;31:1246-1254.
57. Barker Charlotte IS, Snape MD. Pandemic influenza A H1N1 vaccines and narcolepsy: vaccine safety surveillance in action. *Lancet Infect Dis*. 2014;14:227-238.
58. Salmon Daniel A, Michael P, Richard F, et al. Association between Guillain-Barré syndrome and influenza A (H1N1) 2009 monovalent inactivated vaccines in the USA: a meta-analysis. *Lancet*. 2013;381:1461-1468.
59. Clémence G, Pauline B, Jérémie R, et al. Seasonal influenza vaccine and Guillain-Barré syndrome: a self-controlled case series study. *Neurology*. 2020;94:e2168-e2179.
60. Faaizah A, Schuemie Martijn J, Fan B, et al. Serially Combining Epidemiological Designs Does Not Improve Overall Signal Detection in Vaccine Safety Surveillance. *Drug Saf*. 2023;797–807:1-11.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bu F, Schuemie MJ, Nishimura A, et al. Bayesian safety surveillance with adaptive bias correction. *Statistics in Medicine*. 2024;43(2):395-418. doi: 10.1002/sim.9968