Title

Serially Combining Epidemiological Designs Does Not Improve Overall Signal Detection in Vaccine Safety Surveillance.

Permalink

https://escholarship.org/uc/item/5vs3x5qb

Journal

Drug Safety, 46(8)

Authors

Arshad, Faaizah
Schuemie, Martijn
Bu, Fan
et al.

Publication Date

2023-08-01

DOI

10.1007/s40264-023-01324-1

Drug Safety (2023) 46:797–807
https://doi.org/10.1007/s40264-023-01324-1

ORIGINAL RESEARCH ARTICLE

Serially Combining Epidemiological Designs Does Not Improve Overall Signal Detection in Vaccine Safety Surveillance

Faaizah Arshad[1,2] · Martijn J. Schuemie[1,2,3] · Fan Bu[1,2] · Evan P. Minty[4] · Thamir M. Alshammari[5] · Lana Y. H. Lai[6] · Talita Duarte-Salles[7] · Stephen Fortin[3] · Fredrik Nyberg[8] · Patrick B. Ryan[2,3] · George Hripcsak[2,9,10] · Daniel Prieto-Alhambra[11,12] · Marc A. Suchard[1,2,13,14]

Accepted: 29 May 2023 / Published online: 16 June 2023

Abstract

Introduction Vaccine safety surveillance commonly includes a serial testing approach with a sensitive method for 'signal generation' and specific method for 'signal validation.' The extent to which serial testing in real-world studies improves or hinders overall performance in terms of sensitivity and specificity remains unknown.
Methods We assessed the overall performance of serial testing using three administrative claims and one electronic health record database. We compared type I and II errors before and after empirical calibration for historical comparator, self-controlled case series (SCCS), and the serial combination of those designs against six vaccine exposure groups with 93 negative control and 279 imputed positive control outcomes.
Results The historical comparator design mostly had fewer type II errors than SCCS. SCCS had fewer type I errors than the historical comparator. Before empirical calibration, the serial combination increased specificity and decreased sensitivity. Type II errors mostly exceeded 50%. After empirical calibration, type I errors returned to nominal; sensitivity was lowest when the methods were combined.
Conclusion While serial combination produced fewer false-positive signals compared with the most specific method, it generated more false-negative signals compared with the most sensitive method. Using a historical comparator design followed by an SCCS analysis yielded decreased sensitivity in evaluating safety signals relative to a one-stage SCCS approach. While the current use of serial testing in vaccine surveillance may provide a practical paradigm for signal identification and triage, single epidemiological designs should be explored as valuable approaches to detecting signals.

Key Points

Using the serial approach in vaccine safety surveillance did not improve overall performance: specificity increased but sensitivity decreased.

Without empirical calibration, type II errors exceeded 50%; after empirical calibration, type I error rates returned to nominal with negligible change to type II error rates.

While prior research has suggested high sensitivity of the historical comparator method in distinguishing true safety signals, there were cases where a self-controlled case series was more sensitive.

1 Introduction

As of October 2022, over 12 billion doses of coronavirus disease 2019 (COVID-19) vaccines have been administered worldwide [1, 2], with 6200 million in the United States (US) [3, 4] and 600 million in Europe [5]. While the rapid, global increase in vaccine uptake brings promise that the pandemic will end soon, undesirable and sometimes life-threatening adverse events can occur after immunization. Public health experts are employing post-vaccine safety surveillance to identify risk factors and complications of vaccine use.

Typically, in vaccine development, companies in the US are only permitted to market their vaccine if approved by the US Food and Drug Administration (FDA). However, due to the COVID-19 emergency, the US FDA granted emergency use authorizations to permit vaccine distribution ahead of the typical approval process. This

Extended author information available on the last page of the article

underscores the need for post-vaccine monitoring to ensure that unapproved vaccines in widespread use are safe. Often, only a limited number of people participate in clinical trials (which are frequently highly restricted, excluding groups such as pregnant women or persons with specific comorbidities like cancer, etc.), so initial data on rare adverse events can be limited and ungeneralizable to real-world populations [6]. After the vaccine is marketed and provided to larger and more diverse populations, as is the case with the COVID-19 vaccines, previously undetected signals (possible associations between vaccines and adverse events that warrant further investigation) may become apparent [7].

The detection of such signals has traditionally been done via passive surveillance (spontaneous reporting systems) because they can be inexpensive [8]. In the case of observational studies, they can also help expose rare adverse events that occur with short-term or initial medication use [9]. Some downsides, however, are that they are limited by underreporting, selective reporting, and lack of a clear denominator [8]. A recently emerging alternative for generating vaccine safety evidence is through the use of electronic health records (EHRs), claims, and other available real-world clinical data. This approach has been divided into signal generation and signal evaluation [10], which are conducted using epidemiological designs such as historical comparator and self-controlled case series (SCCS). Whereas signal generation uses a highly sensitive but not specific method that can quickly generate hypotheses about signals, signal evaluation involves a complex and computationally expensive method that is highly specific but not sensitive [11]. Monitoring programs that use a two-stage approach include the Vaccine Safety Datalink and the FDA Center for Medicare and Medicaid Services (CMS) [12, 13]; both also suggest the advantages and disadvantages for human expertise to modify the second stage after the data have already been observed.

A widely understood incentive for using a serial approach in vaccine safety surveillance is that it utilizes an inexpensive approach before an expensive one, making it potentially favorable when resources are limited. This expense arises from a requirement for much richer data collection and analysis tools for potentially improved confounding adjustment. A common clinical intuition is that applying a sensitive diagnostic first and following up with a specific test may yield acceptable operating characteristics when two epidemiological designs are applied serially. As new software and analytical techniques are rapidly developed to investigate vaccine safety signals, and the cost of different approaches may decrease and become more similar, it is worth determining the quantitative benefit to using serial signal detection relative to applying single methods alone. A general notion is that sensitivity and specificity have trade-offs such that improvement in one dimension can lead to losses in the other. This study is an opportunity to evaluate the magnitude of loss or gain with serial testing in real-world settings. The extent to which sensitivity and specificity changes has not previously been well appreciated. Contextualizing this trade-off for vaccine surveillance allows us to further improve epidemiological methods. Our purpose is not to advocate for or against the use of serial testing, but to quantify overall sensitivity and specificity when a sensitive design is followed by a specific design. We aim to use real-world evidence to empirically present advantages and drawbacks of the serial testing paradigm.

In this study, we quantified the extent to which the serial strategy (first applying a sensitive method for signal generation and then a specific method for signal evaluation) is useful for signal detection. Epidemiologists commonly use this method for exposure and outcome classification [11]. A ‘gold standard’ serial testing method would have 100% sensitivity and also very high specificity [14]. Prior studies have analyzed the statistical accuracy of combining diagnostic tests [15, 16]; however, they are focused on understanding tests for patient-level screening, rather than population-level signals for surveillance, which are of interest here. We aim to empirically quantify the magnitude to which sensitivity and specificity change by graphically illustrating impact on numerical values for type I and type II error. Understanding of signal detection for surveillance is still recent and not widespread. Given that computational abilities have improved, we herein evaluate serial testing and report on its quantitative performance. This provides valuable data and information to make statements about methods. Monitoring the safety of vaccines is of increasing importance, especially amidst the COVID-19 pandemic which has resulted in unprecedented levels of hospitalization and death, with rigorous efforts to distribute vaccines. Ensuring public safety means carefully evaluating scientific evidence to ensure that epidemiological methods are well grounded. Our research comprehensively addresses this existing gap and has social impact.

It remains unknown the extent to which a serial strategy in vaccine surveillance based on available data improves or reduces the overall sensitivity or specificity of signal detection and validation, and therefore generates fewer false-positive or false-negative population-level signals. In the context of our study design, serial testing implies retesting signals flagged in the historical comparator design using SCCS methods. Specifically, we evaluated performance metrics when the historical comparator design (shown elsewhere to be relatively sensitive) was followed by SCCS (shown elsewhere to be relatively specific) [17], using negative and imputed positive control outcomes in patients receiving historic vaccines across four national administrative claims and EHR databases [18].

2 Materials and Methods

2.1 Databases

To evaluate the comparative performance of epidemiological designs, we obtained de-identified clinical records from four large administrative claims and EHR databases that used the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) in the United States: IBM MarketScan Commercial Claims and Encounters (CCAE), IBM MarketScan Medicare Supplemental Database (MDCR), IBM MarketScan Multi-State Medicaid Database (MDCD), and Optum® de-identified Electronic Health Record dataset (Optum EHR).

Of the four sources, three (IBM CCAE, IBM MDCR, and IBM MDCD) included adjudicated health insurance claims for commercially insured individuals younger than 65 years old, commercially insured individuals 65 years or older, and racially diverse Medicaid enrollees, respectively. Optum EHR contained electronic health records covering the general US population.

Our use of real-world data is important because it represents the general population, where individuals with various health conditions and baseline characteristics are included, unlike controlled studies where samples are homogenous because certain patients are excluded.

The minimum observation period dates for IBM CCAE, IBM MDCR, IBM MDCD, and Optum EHR were January 1, 2000, January 1, 2000, January 1, 2006, and January 1, 2007, respectively. The maximum observation period dates for IBM CCAE, IBM MDCR, IBM MDCD, and Optum EHR were October 31, 2020, October 31, 2020, December 31, 2019, and June 29, 2021, respectively.

Please refer to Table 1 for additional details.

2.2 Exposures of Interest

We selected six historic vaccine exposures of interest: hemagglutinin type 1/neuraminidase type 1 (2009 pandemic influenza) (H1N1pdm), seasonal influenza (Fluvirin), seasonal influenza (Fluzone), seasonal influenza (All), zoster (Shingrix; first or second dose), and human papillomavirus (HPV) (Gardasil 9; first or second dose). Each exposure was evaluated over a specific, historical, 1-year start and end date (see Table 2).

2.3 Negative Control Outcomes

We generated a single list of 93 negative control outcomes (see Supplementary Table 1 in the electronic supplementary material [ESM]), which were outcomes not believed to be caused by any vaccines or outcomes of interest. To identify negative control outcomes that match the severity and prevalence of suspected vaccine adverse effects, we generated a candidate list of negative controls based on similarity of prevalence and percent of diagnoses that were recorded in an inpatient setting (as a proxy for severity). Clinicians reviewed the list to confirm their negative status. Effect size estimates for negative control outcomes should be close to the null; a detected signal away from the null would indicate a type I error.

2.4 Positive Control Outcomes

We modified the 93 negative controls to generate 279 imputed positive control outcomes (outcomes known to be caused by the vaccines) by multiplying the estimated effect size of each negative control by 1.5, 2, and 4. Here's an example. After the design has been applied, assume the point estimate for a negative control outcome is 1.03. To generate an imputed positive control outcome that has a true rate ratio of 1.5, we multiplied the point estimate 1.03 by a factor of 1.5. The estimate for the imputed positive control (with true rate ratio = 1.5) would be 1.03 × 1.5 = 1.545. We chose imputed positive controls instead of real positive controls to evaluate type II errors because well-established vaccine adverse events are rare or carefully monitored such that real-world data do not clearly convey the risk and magnitude of their associations. These simulated positive controls retain the same complex biases and confounding that exist in the negative controls from which they are imputed. We recognize that these controls may not protect against different forms of biases not present in the negative controls since, for example, they assume that systematic error does not change as a function of true effect size.

2.5 Choice of Epidemiological Designs

In a previous study by Schuemie et al., we evaluated eight variations of the historical comparator cohort design and five variations of the self-controlled case series. We found that using different methods on the same data can result in heterogeneous estimates. For example, some methods overestimated the effect size. Detailed analysis can be found in the ESM [17]. To assess the hypothesis that serial testing is favorable, we chose one epidemiological design suspected to be highly sensitive but not highly specific (historical comparator) [17]. The historical comparator design compares the observed incidence of adverse events following immunization (AEFI) with the expected incidence of AEFI, estimated from an unexposed patient population, often before or 'historical' to vaccine introduction. We adjusted for age and sex, as recommended by a previous study [19], and used a 'time at risk' of 1–28 days after the historic visit (defined as a

Table 1 Three administrative claims and one EHR data sources with details on population, patients, and history

Data source	Population	Patients	History	Description
IBM MarketScan Commercial Claims and Encounters Database (CCAE)	Commercially insured, <65 years	142M	2000–	Enrollees in US employer-sponsored insurance health plans. Adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy), lab tests, and enrollment data from large employers and health plans who provide private healthcare coverage. Includes various fee-for-service, preferred provider organizations, and capitated health plans
IBM Health MarketScan® Multi-State Medicaid Database (MDCD)	Medicaid enrollees, racially diverse	26M	2006–	Adjudicated US health insurance claims for Medicaid enrollees. Contains hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims. Lacks lab data
IBM MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR)	Commercially insured, 65+ years	10M	2000–	Health services of retirees in US with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. Adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy). Lab tests for some
Optum de-identified Electronic Health Record Dataset (Optum EHR)	US, general	93M	2006–	Derived from dozens of healthcare provider organizations in US (700+ hospitals and 7000+ clinics treating 103+ million patients). Clinical information, prescriptions, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical notes using Natural Language Processing

Table 2 Six historic (groups of) vaccines evaluated over specific 1-year start and end dates

Exposure name	Shot	Start date	End date	History start date	History end date
H1N1 vaccination	First	Sep 1, 2009	May 31, 2010	Sep 1, 2008	May 31, 2009
Seasonal flu vaccination (Fluvirin)	First	Sep 1, 2017	May 31, 2018	Sep 1, 2016	May 31, 2017
Seasonal flu vaccination (Fluzone)	First	Sep 1, 2017	May 31, 2018	Sep 1, 2016	May 31, 2017
Seasonal flu vaccination (All)	First	Sep 1, 2017	May 31, 2018	Sep 1, 2016	May 31, 2017
Zoster vaccination (Shingrix)	Both	Jan 1, 2018	Dec 31, 2018	Jan 1, 2017	Dec 31, 2017
HPV vaccination (Gardasil 9)	Both	Jan 1, 2018	Dec 31, 2018	Jan 1, 2017	Dec 31, 2017

Refer to Appendix A of the study protocol for the formal cohort definitions of each exposure

random outpatient visit) for both first and second doses. Any outcomes during that time were attributed to the vaccine [19]. Early signal detection during the historical comparator method can be done if some adverse events are present; however, this can also introduce confounding or false/missed signaling if background rates are inexact or change over time [20]. Then, we selected one design that was suspected to be specific but not sensitive (SCCS) [17]. SCCS compares the time shortly following vaccination to all other time in the same patient's record, therefore focusing our study on immediate or short-term adverse events [21, 22]. We used a SCCS design adjusted for age and season and excluded a 30-day pre-vaccination window from the analyses to account for healthy vaccinee bias [23].

2.6 Performance Metrics

We computed effect size estimates with 95% confidence intervals (CIs) and one-sided *p*-values across all databases and all exposures for the historical comparator method alone, SCCS method alone, and the methods combined (historical comparator followed by SCCS). We distinguished signals from non-signals by using $p < 0.05$ when both designs were applied separately, as well as when they were applied serially (with SCCS applied to signals generated by the historical comparator).

A primary concern in observational studies is the presence of systematic errors that may exist because exposed and unexposed populations are not experimentally randomized, but simply observed [24]. Unlike random errors, systematic errors do not approach zero by merely increasing sample size [25]. It is therefore relatively more problematic when using large databases, such as in our study, which capture records for millions of patients. Empirical calibration is a statistical procedure used to adjust for such systematic errors [26]. It derives a null distribution from a sample of negative control outcomes and then applies the distribution to unknown effect size estimates, calibrating *p*-values so that 5% of negative controls have $p < 0.05$ [25]. We can similarly calibrate 95% CIs by modifying negative controls to synthesize positive controls. After calibration, the coverage for a 95% CI is closer to the expected 95% [25]. Without calibration, certain biases may be left unaccounted for [26]. One important assumption of empirical calibration is that the systematic error of the exposure–outcome pair of interest draws from the same distribution as the systematic error for the negative control. Even though it does not require any negative control to have the exact same confounding structure as the exposure–outcome of interest, a weaker assumption of exchangeability is still required. Here we use *p*-value calibration and make no assumption about systematic error when the true effect size differs from the null.

We analyzed type I and II errors before and after empirical calibration. A tradeoff exists between the two errors, leading to an increase in type II error depending on how much systematic error is adjusted for.

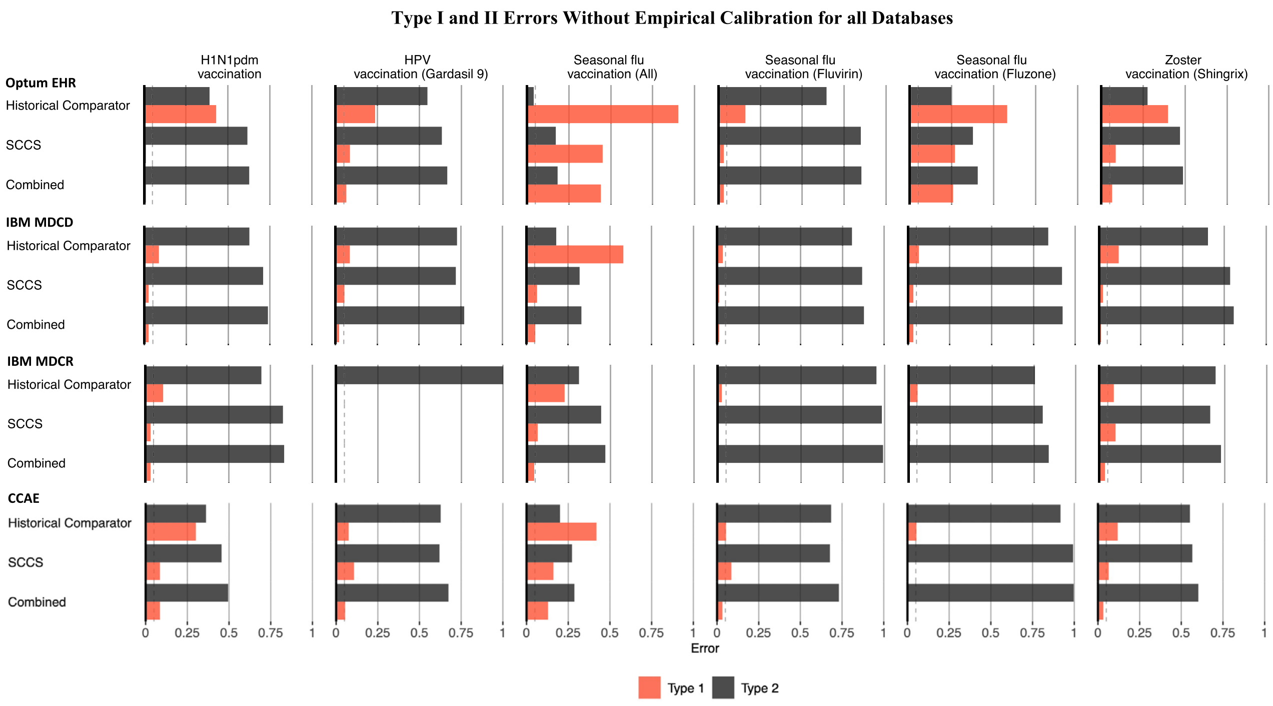
Please refer to https://github.com/ohdsi-studies/Eumaeus to access the protocol and analytic code. The OHDSI self-controlled case series repository can be found at https://github.com/OHDSI/SelfControlledCaseSeries. The OHDSI Cohort Method repository can be found at https://github.com/OHDSI/CohortMethod.

3 Results

We define the following terms to ease comprehension of the findings. Type I error (also known as a false-positive conclusion) is when the system claims a signal when not true. Low type I error indicates high specificity. High type I error indicates low specificity. Type II error (also known as a false-negative conclusion) is when the system misses a signal that exists. Low type II error indicates high sensitivity. High type II error indicates low sensitivity.

3.1 Type I and II Errors Before Empirical Calibration for All Databases

Figure 1 reports type I and II error rates in classifying positive and negative control outcomes across vaccine exposure



Optum EHR = Optum Electronic Health Records; IBM MDCD = IBM Health MarketScan® MultiState Medicaid Database;
IBM MDCR = IBM MarketScan Medicare Supplemental and Coordination of Benefits Database; CCAE = IBM MarketScan Commercial Claims and Encounters Database;
H1N1pdm = hemagglutinin type 1 / neuraminidase type 1 (2009 pandemic influenza); HPV = human papillomavirus.

Fig. 1 We depict uncalibrated type I and II errors across all databases using all exposures and negative and positive controls during a 'time at risk' of 1–28 days following vaccination. Generally, the historical comparator had a low type II error rate. SCCS had a low type I error rate. Clinical intuition would predict the combined method to be highly sensitive and specific, but results showed high specificity and low sensitivity

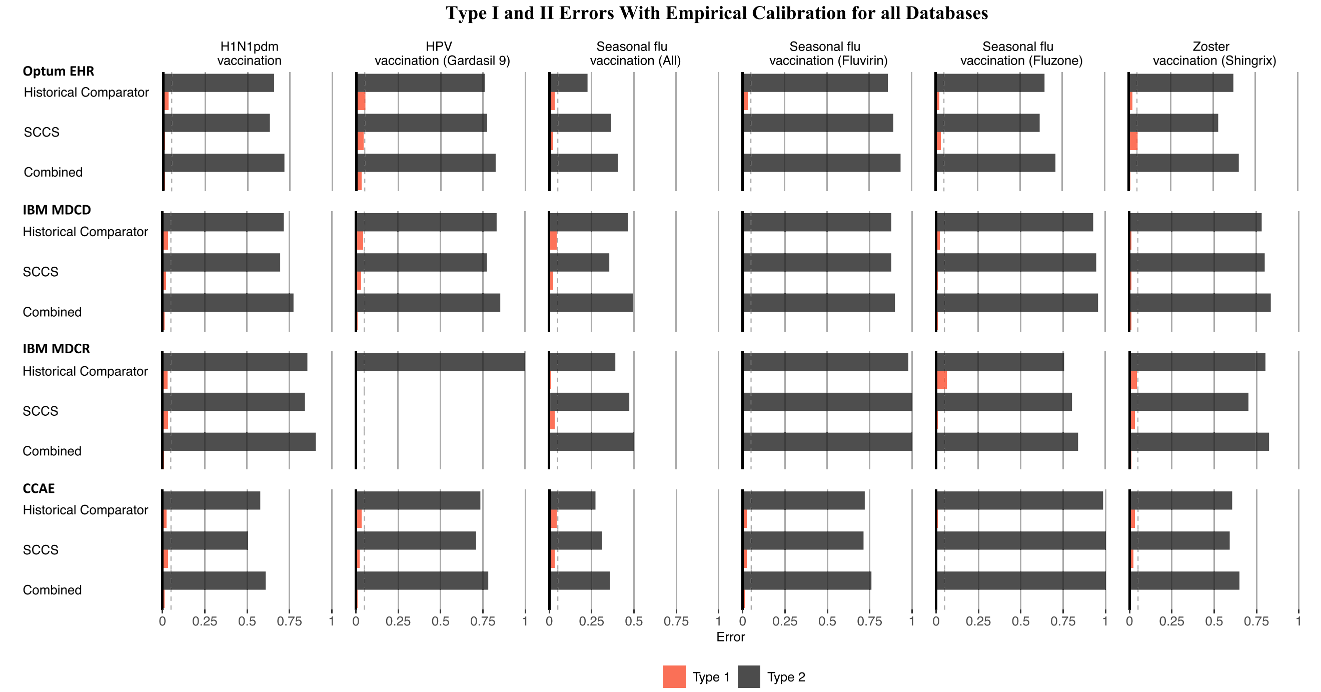
groups. We further compare historical comparator and SCCS designs with the serial combination of those designs on the *y*-axis. Bars in orange indicate type I error; bars in gray indicate type II error. Type I error rate has a nominal cutoff of 0.05 as shown by the dotted line to the right of 0. Error bars closer to 0 indicate higher sensitivity and specificity.

Almost all analyses had type II error rates exceeding 50% (Fig. 1). Meanwhile, type I error bars tended to range between 0 and 50%, except for seasonal influenza (All) in the Optum EHR database where the historical comparator method generated many false-positive signals, and seasonal influenza (Fluzone) in the Optum EHR database where SCCS showed a type 1 error rate slightly exceeding 50%.

The difference between type I and II error rates before empirical calibration was particularly visible for the Optum EHR database. In most scenarios, the historical comparator method had higher sensitivity than SCCS or the combined method. When the historical comparator design was used to identify adverse events following vaccination, it generated fewer false-negative signals (was less likely to miss a signal when it existed). Meanwhile, SCCS was more specific than the historical comparator method for H1N1pdm, seasonal influenza (All), seasonal influenza (Fluvirin), and zoster (Shingrix). The serial approach generated more false-negative signals, while often reducing the number of false positives. This was consistent across all four databases: serial combination did not improve overall performance.

3.2 Type I and II Errors After Empirical Calibration for All Databases

We also considered how empirical calibration changed type I and II errors (Fig. 2). After calibration, type I errors returned to nominal, and type II errors increased in most cases. For the Optum EHR database, calibration visibly reduced type I error rates of the historical comparator method for seasonal influenza (All). In spite of this, the combined method did not improve overall performance in the way that our hypothesis predicted. Among all three design choices and databases, sensitivity was lowest when the methods were serially combined (it increased the number of times that the surveillance system missed a signal when it actually existed). These results may be expected as only signals flagged in the historical comparator method are then evaluated in SCCS, so there is an opportunity to decrease false-positive signals. This makes the hypothesis more relevant because it questions why



Optum EHR = Optum Electronic Health Records; IBM MDCD = IBM Health MarketScan® MultiState Medicaid Database;
IBM MDCR = IBM MarketScan Medicare Supplemental and Coordination of Benefits Database; CCAE = IBM MarketScan Commercial Claims and Encounters Database;
H1N1pdm = hemagglutinin type 1 / neuraminidase type 1 (2009 pandemic influenza); HPV = human papillomavirus.

Fig. 2 We depict type I and II errors after empirical calibration across all databases using all exposures and negative and positive controls during a 'time at risk' of 1–28 days following vaccination. Type I errors returned to nominal. Type II errors mostly increased. Even with empirical calibration, the serial approach did not increase overall performance of a two-stage surveillance program

surveillance systems are implementing a procedure that does not improve overall performance.

4 Discussion

4.1 Key Results

In this large-scale observational study of vaccine surveillance using four real-world databases in the US, we examined the benefit of using a serial testing approach. A common clinical intuition and hypothesis in surveillance is that both sensitivity and specificity improve when two epidemiological designs are applied serially. Contrary to popular belief, there are theoretical arguments against this, and we found empirically that combining designs does not improve overall performance. Although serial combination reduced type I error, it was flawed because type II error increased to be higher than when either of the methods were used alone. It is clear that this tradeoff in type I and II errors is expected. Only signals flagged in the historical comparator were evaluated in the second method, SCCS, creating an opportunity to remove false-positive signals but not false-negative signals. Also, SCCS typically already had quite favorable type I error rates when applied on its own, so combining methods did little to further decrease type I errors. Despite these somewhat predictable results, our motive was to quantify and compare the losses and gains in using one kind of design (either historical comparator or SCCS alone) with using two designs together, to provide empirical guidance for practical safety surveillance. Our results, based on real-world reliable evidence, show that the notion of dividing surveillance into two distinct stages of signal generation and evaluation might not be ideal. False-negative signals were excessively higher than false-positive signals.

Some might argue that our findings were predictable from the start. As noted previously, one can indeed mathematically demonstrate that sensitivity decreases and specificity increases when positives are retested. However, the extent to which sensitivity decreases and specificity increases was previously not well appreciated, which might explain why current epidemiological practices still often utilize the serial approach, despite an indication that overall signal detection performance might not improve. The serial approach also allows human expertise to guide the second design in making non-pre-specified design choices. This, however, can introduce further bias in the sense of steering the analysis based on the observed data and result in p-hacking [27, 28]. Using real-world evidence to generate results allows us to obviate a problem that exists in the current paradigm. The math behind this finding might not be novel, but contextualizing it for vaccine surveillance is important.

The novelty in our study is that we empirically quantified the magnitude to which type I and type II errors changed when each design (historical comparator and SCCS) was applied individually, versus when they were applied together. We also found that using a single design might bring benefits over using two designs. This is not to say that serial testing is not a worthwhile method, or that there is a hierarchy of methods. Rather, it is to suggest that the existing paradigm has its inadequacies and might not achieve the intended type I and type II error outcomes that would be desirable for an ideal surveillance design. In other words, using one design alone as opposed to two together could be valuable. We encourage vaccine monitoring systems to consider the relative gains that could come from using one design alone.

The preferred tradeoff between sensitivity and specificity in safety surveillance is not a given, and may depend on many considerations. Sensitivity alone does not predict enough about negative signals, and specificity alone does not predict enough about positive signals ('worst of both worlds'). If specificity is high, but sensitivity is low, potential adverse events that might be caused by a vaccine can be overlooked, allowing for further circulation of the vaccine in the market. If sensitivity is high and specificity is low, adverse events might be perceived as associated with a vaccine when in truth they are not, leading to the withdrawal of a vaccine from the market. In either case, an error has been made.

Also note that sensitivity and specificity are parameters that vaccine safety surveillance monitoring systems intentionally control to meet performance expectations. So, the results (decreased sensitivity and increased specificity) are not a product of arbitrary mathematical explanation, but a consequence of flawed reasoning (purposely selecting an extra sensitive system for the first epidemiological design and an extra specific system for the second epidemiological design) that could have been due to resource constraints in the past. With better analytical techniques now, our study is an opportunity to identify weaknesses in the existing surveillance paradigm and stimulate conversations that further improve surveillance methods.

Other studies have also examined performance characteristics in serial monitoring. Nelson et al. studied the use of serial approaches with observational EHR data in identifying signals between the DTap-IPV-Hib vaccine and adverse events [29]. One analytic challenge was confounding introduced from differences in vaccine uptake by age group [29]. Since observational studies are inherently uncontrolled, we corrected for confounding via empirical calibration.

Prior research has suggested that the historical comparator method is highly sensitive. For example, Li et al. recently found that historical comparisons had type II errors ranging between 0 and 10% and type I errors above 30% [19]. Age and sex adjustment reduced type I errors, had little impact on type II errors, and improved precision in some cases [19]. Historical rate comparison was, overall, good at identifying true safety signals [19]. Based on Li et al.'s results, we evaluated the age- and sex-adjusted variant for historical comparator, using a 'time at risk' of 1–28 days after the historic visit, for both first and second doses. However, contrary to this theory, we observed some instances where SCCS had lower type II error rates and was therefore more sensitive than the historical comparator method. For example, in Fig. 1, this was present for outcomes involving HPV (Gardasil 9) and zoster (Shingrix) in the Optum EHR and CCAE databases. There was also almost no distinguishable difference between the type II errors for the historical comparator and SCCS methods in several of the plots, including for H1N1pdm and seasonal influenza (Fluvirin) in IBM MDCD, as well as seasonal influenza (Fluzone) in IBM MDCR. Likewise, there were some incidences where SCCS was less specific than the historical comparator method, such as in data for seasonal influenza (Fluzone) in Optum EHR, zoster (Shingrix) in IBM MDCR, and HPV (Gardasil 9) in CCAE. Comparisons of performance metrics demonstrate that the historical comparator is not always most sensitive. This uncertainty can make it difficult to predict the utility of a serial approach. Finally, Maro et al. conducted a similar study on sequential database surveillance using first an unmatched historical comparison method and second a self-controlled risk interval design (SCRI) [30]. Their 'time at risk' was 1–42 days. However, this study implemented both methods using a sequential framework instead of a serial approach and provided limited information on the sensitivity and specificity of combining methods. The primary difference is that a sequential test repeatedly tests for associations between drugs and adverse events as data accumulates, so test statistics are taken on slightly different datasets. However, serial testing applies the same designs to the same data.

Our results might be relevant and applicable not only at the population level but also at the individual patient-level scale. The serial approach is common worldwide in the realm of diagnostic testing. Consider, for instance, detecting human immunodeficiency virus (HIV) prevalence in a population. A study amongst Nigerian women found that serial rapid tests (where the first HIV test has a sensitivity close to 100%, and the second test is implemented only if the initial results were positive) were more sensitive and specific than conventional HIV testing techniques [31]. Similar studies that used HIV rapid serological testing using data from Cote d'Ivoire [32], Uganda [33], and India [34] found that serial testing provides reliable results and has benefits in resource-limited areas. Serial testing has been widely adopted in clinical screenings for tuberculosis [35], prostate cancer [36], acute coronary syndrome [37], ovarian cancer [38], and recently COVID-19 [39].

This principle of investigating diseases using a cheap method before an expensive one is central to clinical medicine and could be worth studying further. The quantitative losses that come with using serial testing might be worth facing if the cost saving and convenience are important factors in the decision for surveillance methodology.

The importance of our study for the future of safety surveillance is related to recommended measures that should be addressed before using observational electronic healthcare databases for safety surveillance [40], namely that of prespecifying and evaluating statistical designs. An accurate epidemiological design should overall generate the fewest number of false-positive and false-negative signals. Combining designs is unlikely to achieve this goal.

Individuals who prioritize specificity over sensitivity might claim serial combination did work. Further studies should examine when this tradeoff is worthwhile. Since we observed variability in results, there should be similar analyses with other databases and exposures, such as COVID-19 vaccines or populations for which a certain occurrence has been defined (e.g. pregnant women or those with common comorbidities).

4.2 Strengths and Limitations

Three strengths of this work are the use of real-world data, mapping of data to the OMOP CDM, and open-source availability of our code. First, real-world data represents a more heterogeneous population than randomized controlled trials that are tightly controlled. This improves the external validity of our data to the real-world patient population. Second, the CDM unifies data from multiple partners into a standardized format and enables us to pool analysis results from sources with large numbers of study subjects, without having to expose patient-level information. Third, the open-source availability of our code and protocol enables transparency and collaboration.

Our study is limited by the vaccine exposure of interest. While we seek to understand the use of epidemiological designs not only in general but also for COVID-19 vaccine surveillance, none of our exposures were COVID-19 vaccines. We addressed this limitation by choosing historical viral vaccines, which may be most similar. Another limitation is the use of imputed positive controls that assumes that systematic error does not change as a function of true effect size.

5 Conclusion

As COVID-19 vaccine safety surveillance becomes increasingly important, monitoring systems should carefully consider the utility and sequence of epidemiological designs. Sensitivity and specificity of surveillance methods plays an important role in how well the system identifies signals. While serial combination produced fewer false-positive signals compared with the most specific method, it generated more false-negative signals compared with the most sensitive method. Using a historical comparator design followed by an SCCS analysis yielded decreased overall performance in evaluating safety signals relative to a one-stage approach. While the current use of serial testing in vaccine surveillance may provide a practical paradigm for signal identification and triage, single epidemiological designs should be explored as valuable approaches to detecting signals.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-023-01324-1.

Acknowledgements We are indebted to the entire Evaluating Use of Methods for Adverse Event Under Surveillance for vaccines (EUMAEUS) task force within the Observational Health Data Sciences and Informatics community.

Declarations

Funding This work was supported by US National Institutes of Health [grant number R01AI153044], a contract from the US Food and Drug Administration and an Intergovernmental Personnel Act agreement with the US Department of Veterans Affairs.

Conflict of Interest FA and FB declare no competing interests. MAS receives grant funding from the US National Institutes of Health and the US Food and Drug Administration and contracts from the US Department of Veterans Affairs and Janssen Research and Development. PBR, SF and MJS are employees of Janssen Research and Development and shareholders in Johnson & Johnson. GH receives grant funding from the US National Institutes of Health and the US Food and Drug Administration. DPA's research group has received grant support from Amgen, Chesi-Taylor, Novartis, and UCB Biopharma. His department has received advisory or consultancy fees from Amgen, Astellas, AstraZeneca, Johnson, and Johnson, and UCB Biopharma and fees for speaker services from Amgen and UCB Biopharma. Janssen, on behalf of IMI-funded EHDEN and EMIF consortiums, and Synapse Management Partners have supported training programmes organized by DPA's department and open for external participants organized by his department outside submitted work. FN was an employee of AstraZeneca until 2019 and owns some AstraZeneca shares. EPM, LYHL, TMA, and TDS declare no competing interests.

Ethics Approval This study did not require Institutional Review Board approval from data partners because it used de-identified patient data and did not constitute research with human subjects.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Data Availability The data supporting the findings of this study are available at https://data.ohdsi.org/Eumaeus/.

Code Availability The protocol and analytical code underlying this manuscript are available in the Eumaeus Repository at https://github.com/ohdsi-studies/Eumaeus.

Author Contributions FA, MJS, PBR, GH, DPA and MAS conceived the research. FA, MJS and MAS drafted the manuscript in consultation with EPM, TMA, LYHL, TDS, SF, FN, FB, PBR, GH, and DPA, who provided critical feedback on the research and writing. All authors read and approved the final version.

References

1. Bloomberg [Internet]. https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/ [cited 18 Nov 2021].
2. WHO Coronavirus (COVID-19) dashboard [Internet]. https://covid19.who.int [cited 14 Dec 2022].
3. CDC. COVID Data Tracker [Internet]; 2020. https://covid.cdc.gov/covid-data-tracker/ [cited 28 July 2021].
4. Holder J. Tracking Coronavirus Vaccinations Around the World. The New York Times [Internet]; 2021 Jan 29. https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html [cited 16 Aug 2021].
5. COVID-19 Vaccine Tracker [Internet]. https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html [cited 28 July 2021].
6. Kim HS, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. J Korean Med Sci. 2018;33(34): e213.
7. [Internet]. https://cioms.ch/wp-content/uploads/2018/03/WG8-Signal-Detection.pdf [cited 1 May 2023].
8. Kumar A, Khan H. Signal detection and their assessment in pharmacovigilance. Open Pharm Sci J. 2015;2(1):66–73.
9. Li Y, Ryan PB, Wei Y, Friedman C. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. Drug Saf. 2015;38(10):895.
10. Mesfin YM, Cheng A, Lawrie J, Buttery J. Use of routinely collected electronic healthcare data for postlicensure vaccine safety signal detection: a systematic review. BMJ Glob Health. 2019;4(4): e001065.
11. https://www.brookings.edu/wp-content/uploads/2012/04/Setting-Priorities_Discussion-guide_20110527.pdf [cited 1 May 2023].
12. Vaccine Safety Datalink (VSD) [Internet]; 2022. https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vsd/index.html [cited 1 May 2023].
13. [Internet]. https://www.fda.gov/media/71238/download [cited 1 May 2023].
14. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhal Toxicol. 2014;26(13):811–28.
15. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. Biostatistics. 2000;1(2):123–40.
16. Thompson ML. Assessing the diagnostic accuracy of a sequence of tests. Biostatistics. 2003;4(3):341–51.
17. Schuemie MJ, Arshad F, Pratt N, Nyberg F, Alshammari TM, Hripcsak G, et al. Vaccine safety surveillance using routinely collected healthcare data-an empirical evaluation of epidemiological designs. Front Pharmacol. 2022;6(13): 893484.
18. Protocol.utf8 [Internet]. https://ohdsi-studies.github.io/Eumaeus/Protocol.html [cited 30 Sep 2021].
19. Li X, Lai LY, Ostropolets A, Arshad F, Tan EH, Casajust P, et al. Bias, precision and timeliness of historical (background) rate comparison methods for vaccine safety monitoring: an empirical multi-database analysis. Front Pharmacol. 2021;24(12): 773875.
20. Yih WK, Kulldorff M, Fireman BH, Shui IM, Lewis EM, Klein NP, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics. 2011;127(Suppl 1):S54–64.
21. Whitaker HJ, Paddy Farrington C, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method [Internet]. Stat Med. 2006;25:1768–97. https://doi.org/10.1002/sim.2302.
22. Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs [Internet]. BMJ. 2016;354: i4515. https://doi.org/10.1136/bmj.i4515.
23. Weldeselassie YG, Whitaker HJ, Farrington CP. Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. Epidemiol Infect. 2011;139(12):1805–17.
24. 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(2):209–18.
25. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci USA. 2018;115(11):2571–7.
26. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. Stat Med. 2016;35(22):3883–8.
27. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8): e124.
28. Schuemie MJ, Ryan PB, Pratt N, Chen R, You SC, Krumholz HM, et al. Principles of large-scale evidence generation and evaluation across a network of databases (LEGEND). J Am Med Inform Assoc. 2020;27(8):1331–7.
29. Nelson JC, Yu O, Dominguez-Islas CP, Cook AJ, Peterson D, Greene SK, et al. Adapting group sequential methods to observational postlicensure vaccine safety surveillance: results of a pentavalent combination DTaP-IPV-Hib vaccine safety study. Am J Epidemiol. 2013;177(2):131–41.
30. Maro JC, Brown JS, Dal Pan GJ, Kulldorff M. Minimizing signal detection time in postmarket sequential analysis: balancing positive predictive value and sensitivity. Pharmacoepidemiol Drug Saf. 2014;23(8):839–48.
31. Mbachu II, Udigwe G, Joseph I, John O, Samuel UO, Joseph U, et al. The evaluation of accuracy of serial rapid HIV test algorithm in the diagnosis of HIV antibodies among pregnant women in south east Nigeria. BMC Res Notes. 2015;12(8):557.
32. Koblavi-Dème S, Maurice C, Yavo D, Sibailly TS, N'guessan K, Kamelan-Tano Y, et al. Sensitivity and specificity of human immunodeficiency virus rapid serologic assays and testing algorithms in an antenatal clinic in Abidjan Ivory Coast. J Clin Microbiol. 2001;39(5):1808–12.
33. Galiwango RM, Musoke R, Lubyayi L, Ssekubugu R, Kalibbala S, Ssekweyama V, et al. Evaluation of current rapid HIV test algorithms in Rakai, Uganda. J Virol Methods. 2013;192(1–2):25–7.
34. Iqbal HS, Solomon S, Murugavel KG, Solomon SS, Balakrishnan P. Evaluation of two indigenous rapid and two ELISA assays for the diagnosis of HIV infection India. Indian J Med Microbiol

[Internet]. 2012;30(4). https://pubmed.ncbi.nlm.nih.gov/23183462/ [cited 17 Dec 2021].

35. Moses MW, Zwerling A, Cattamanchi A, Denkinger CM, Banaei N, Kik SV, et al. Serial testing for latent tuberculosis using QuantiFERON-TB Gold In-Tube: a Markov model. Sci Rep. 2016;6(1):1–8.

36. Smith DS, Catalona WJ, Herschman JD. Longitudinal screening for prostate cancer with prostate-specific antigen. JAMA. 1996;276(16):1309–15.

37. Zhelev Z, Ohtake H, Iwata M, Terasawa T, Rogers M, Peters JL, et al. Diagnostic accuracy of contemporary and high-sensitivity cardiac troponin assays used in serial testing, versus single-sample testing as a comparator, to triage patients suspected of acute non-ST-segment elevation myocardial infarction: a systematic review protocol. BMJ Open. 2019;9(3): e026012.

38. Berek JS, Bast RC Jr. Ovarian cancer screening. The use of serial complementary tumor markers to improve sensitivity and specificity for early detection. Cancer. 1995;76(10):2092–6.

39. Kobayashi T, Trannel A, Holley SA, Alsuhaibani MA, Abosi OJ, Jenn KE, et al. Coronavirus disease 2019 serial testing among hospitalized patients in a Midwest Tertiary Medical Center, July–September 2020. Clin Infect Dis. 2020;73(9): e3116–9.

40. Nelson JC, Wellman R, Yu O, Cook AJ, Maro JC, Ouellet-Hellstrom R, et al. A synthesis of current surveillance planning methods for the sequential monitoring of drug and vaccine adverse effects using electronic health care data. EGEMS (Wash DC). 2016;4(1):1219.

Authors and Affiliations

Faaizah Arshad[1,2] · Martijn J. Schuemie[1,2,3] · Fan Bu[1,2] · Evan P. Minty[4] · Thamir M. Alshammari[5] · Lana Y. H. Lai[6] · Talita Duarte-Salles[7] · Stephen Fortin[3] · Fredrik Nyberg[8] · Patrick B. Ryan[2,3] · George Hripcsak[2,9,10] · Daniel Prieto-Alhambra[11,12] · Marc A. Suchard[1,2,13,14]

✉ Marc A. Suchard
msuchard@ucla.edu

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

[2] Observational Health Data Sciences and Informatics, New York, NY, USA

[3] Observational Health Data Analytics, Janssen R&D, Titusville, NJ, USA

[4] O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Calgary, AB, Canada

[5] Medication Safety Research Chair, King Saud University, Riyadh, Saudi Arabia

[6] Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK

[7] Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

[8] School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Göteborg, Sweden

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

[10] Medical Informatics Services, New York-Presbyterian Hospital, New York, NY, USA

[11] Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK

[12] Health Data Sciences, Medical Informatics, Erasmus Medical Center University, Rotterdam, The Netherlands

[13] Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA

[14] VA Informatics and Computing Infrastructure, US Department of Veterans Affairs, Salt Lake City, UT, USA