Received: 13 June 2017 | Revised: 21 October 2017 | Accepted: 23 October 2017

DOI: 10.1002/pds.4356

WILEY

BRIEF REPORT

# Do delays in data availability limit the implementation of near real-time vaccine safety surveillance using the Clinical Practice Research Datalink?

Andreia Leite[1] | Sara L. Thomas[1] | Nick J. Andrews[2]

[1] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2] Statistics, Modelling and Economics Department, Public Health England, London, UK

**Correspondence**
A. Leite, Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.
Email: andreia.leite@lshtm.ac.uk

**Funding information**
National Institute for Health Research

## Abstract

**Purpose:** Near real-time vaccine safety surveillance (NRTVSS) using electronic health records has been used to detect timely vaccine safety signals. Trial implementation of NRTVSS using the Clinical Practice Research Datalink (CPRD) has shown that there is limited power to detect safety signals for rare events. Delays in recording outcomes and receiving data influence the power and timeliness to identify a signal. Our work aimed to compare how different sources of delays influence power and expected time to signal to implement NRTVSS using CPRD.

**Methods:** We studied seasonal influenza vaccine/Guillain-Barré syndrome and performed power and expected time to signal calculations for the 2013-2014/2014-2015 seasons. We used the Poisson-based maximised sequential probability ratio test, which compares observed-to-expected events. For each study season, we obtained an average Guillain-Barré syndrome/seizures age-sex–adjusted rate from the 5 previous seasons and then used this rate to calculate the expected number of events, assuming a 42-day risk-window. Calculations were performed for detecting rate ratios of 1.5 to 10. We compared power and timeliness considering combinations of the presence/absence of delays in recording outcomes and in receiving data. The R-package Sequential was used.

**Results:** In general, there was ≥80% power to detect increases in risk of ≥4 at the end of the season. Assuming absence of delays slightly improved power (a maximum increase of 4%) but did not noticeably reduce time to detect a signal.

**Conclusion:** Removing delays in data availability is insufficient to significantly improve the performance of a NRTVSS system using CPRD. Expansion of CPRD data is required.

KEYWORDS

delay, electronic health records, pharmacoepidemiology, power, safety, surveillance, vaccines

## 1 | INTRODUCTION

Near real-time vaccine safety surveillance (NRTVSS) is an option in the post-licensure vaccine safety toolkit. Near real-time vaccine safety surveillance is usually initiated soon after a new vaccine is introduced, and data from electronic heath records are examined at regular points in time. This helps with timely detection of safety signals.[1]

Near real-time vaccine safety surveillance has not been fully implemented in the UK, but our recent study trialling NRTVSS implementation using data from the Clinical Practice Research Datalink (CPRD) showed it is possible to implement a system.[2] Nevertheless,

This work has not been submitted or accepted elsewhere. Preliminary results have been presented at the 2017 International Society for Pharmacoepidemiology mid-year meeting, London, April, 2017.

system performance (including power and expected time to signal) to identify a rare outcome (Guillain-Barré syndrome, GBS) following seasonal influenza was not optimal. In particular, using the most powerful test (Poisson-based maximised probability ratio test, PMaxSPRT), there was power of ≥80% to identify a fourfold increase in risk, and a signal would be detected 3 months after the start of surveillance. It is thus important to understand what factors affect power and expected time to signal and what changes to currently available data might improve the ability to identify signals rapidly using CPRD.

If PMaxSPRT is used, the expected number of events at the end of the surveillance period dictates power and expected time to signal. The expected number of events is a function of the data available, depending on both the number of individuals contributing data (the volume of data) and on delays in data availability. Clinical Practice Research Datalink is a primary care database, and the volume of data is determined by the number of practices and patients contributing data. Delays can occur in (i) identifying a condition after the initial consultation, (ii) recording a condition diagnosed outside the practice (e.g. in hospital), (iii) practices uploading their data to CPRD, and (iv) researchers receiving data for analysis. Previous work assessing delays due to (ii) showed that, for selected conditions of interest regarding vaccine safety, records accrue within a month of the deemed date of diagnosis.[3] Regarding (iii) and (iv), CPRD data are made available to researchers monthly and practices upload data prior to this, with the last collection date from each practice recorded in CPRD.

Clinical Practice Research Datalink is a dynamic database, and new practices may start contributing data. Additionally, changes to the mode of data collection from practices and frequency of data releases could reduce delays. Both expansion and reduction of delays could improve NRTVSS system performance. We sought to assess how delays influenced power and expected time to signal, to inform data providers on how decreasing delays could improve performance of a NRTVSS system. As a secondary objective, we further assessed the performance of a system based on data available around the middle of the surveillance period for a short vaccine programme of fixed length, to understand what could be detected at a time when it would still be possible to implement measures to minimise risks.

**KEY POINTS**

- The Clinical Practice Research Datalink (CPRD) can be used to implement near real-time vaccine safety surveillance, but there is limited power to detect signals for rare outcomes.
- Delays in recording outcomes and in receiving data might limit power and timeliness of a system. We assessed the influence of these sources of delays to inform data providers of the steps required to improve a system using CPRD data.
- Removing delays in recording outcomes and receiving data is unlikely to significantly improve the performance of a system using CPRD data. Expansion of the data available is needed.

## 2 | METHODS

We used data from our previous study that evaluated the feasibility of implementing a NRTVSS system. Here, we provide a brief summary of the methods used to obtain those data (for further information see Leite et al[2]). Additionally, we explain how we assessed the influence of delays on power and expected time to signal, the main focus of this report.

### 2.1 | Data source

We used CPRD, a UK database containing anonymised primary care records from individuals registered with participating practices (6.9% of the population). Information is Read-coded, including demographics, diagnoses, therapies, vaccinations, health-related behaviours, and referrals to and feedback from hospital.[4] Clinical Practice Research Datalink also contains information of when a patient joined and left a practice (current registration date and transfer out date, respectively), when a practice met certain requirements necessary for it to be considered of research quality (up-to-standard) and when information was last collected from each practice (last collection date, available in each monthly update).[4]

### 2.2 | Vaccine/outcome pairs and study period

Our original study evaluated seasonal influenza vaccine/GBS and mumps-measles-rubella vaccine/seizures. As there was sufficient power to detect a twofold increase in risk for mumps-measles-rubella vaccine/seizures, we considered the performance of the system for this pair was satisfactory. We thus only assessed the influence of delays for seasonal influenza/GBS. We included individuals aged ≥65 years and studied seasons 2013/2014 and 2014/2015, using data released in July 2015 and 2016, respectively.

### 2.3 | Analysis

We used continuous PMaxSPRT as it is the most powerful test, and CPRD provides data in a near-continuous fashion (monthly).[5] The number of expected events was obtained based on the average GBS age-sex–specific rate from the 5 seasons prior to the study seasons (2008-2013 and 2009-2014), considering a 42-day post-vaccination risk-window.

We applied the historical rates to the follow-up time in the study periods to obtain the expected number of events. Start of follow-up time was the latest of the up-to-standard date, current registration date (plus 1 year to exclude retrospective recording of events when registering with a new practice[6]), the beginning of the study period, and the start of the risk-window. End of follow-up was the earliest of the patient's transfer out date, the practice's last collection date, end of the study period, or end of the risk-window.

The number of expected events was calculated in slightly different ways, to consider different delay scenarios (see below). Based on these numbers, we calculated power and expected time to signal (performance measures), assuming a range of plausible rate ratios (1.5-10),

**TABLE 1** Combination of delays assessed under each scenario

| Scenario—Source of Delays | Delays | | End of Surveillance | Comments |
|---|---|---|---|---|
| | Recording | Receiving | | |
| Recording/receiving (reference) | + | + | April data release (end of season) | Corresponds to the way NRTVSS was implemented using CPRD data. Reference scenario |
| 1. None | − | − | April data release (end of season) | Ideal scenario; events are recorded as they happen and data are available immediately |
| 2. Recording | + | − | April data release (end of study period) | Mimics a situation where CPRD receives data on a daily basis and makes it available straight away |
| 3. Recording/receiving | + | + | December data release | Corresponds to the reference scenario but considering data available until December |

Abbreviations: CPRD, Clinical Practice Research Datalink; NRTVSS, near real-time vaccine safety surveillance.

a level of significance of 5%, and stipulating a minimum of 1, 2, or 4 events before raising a signal. Calculations were performed using the R package Sequential.[7]

We assessed the influence of delays on system performance by calculating follow-up time (hence, the expected number of events) assuming the system had different combinations of presence/absence of delays in recording outcomes and in receiving data. Additionally, we looked at performance measures assuming analyses ended at the mid-season (December release). Ending surveillance earlier might increase power as less sequential tests are performed, but the number of expected events is likely to be lower (due to less data available), thus reducing power. The delay scenarios assessed are presented in Table 1. The scenario considering both sources of delays was used as a reference, as this corresponded to what we did for the test implementation.[2]

For delays in recording outcomes, we considered the follow-up time for the patients as explained above (absence of delays) and then adjusted this follow-up time to account for delays, by reducing the expected number of events based on the historical delays' distribution (presence of delays).

For delays in receiving data, we included all data available for the study period regardless of when these data were received (absence of delays) and then included only data received by the end of the surveillance period (presence of delays). We identified data received by the end of surveillance by using the last collection date in that data release. For the reference scenario, we considered the last collection dates available in the April 2014 and April 2015 releases for season 2013/2014 and 2014/2015, respectively. Similarly, we used the last collection dates available in the December releases (2013 and 2014 for season 2013/2014 and 2014/2015, respectively) when assessing performance at the mid-season (scenario 3).

## 3 | RESULTS

Table 2 presents the results of our calculations. In general, there was ≥80% power to detect increases in risk of ≥4 at the end of the season. Removing sources of delays improved power by 1% to 4% and would allow detection of a signal at the same release of the implementation scenario. Stopping surveillance around mid-season (scenario 3) resulted in substantial reductions in power, particularly to detect medium (3-6 fold) increases in risk. For this scenario, there was ≥80% power to detect an increase in risk of 8 to 10. If there was a signal, this would be detected by early December.

## 4 | DISCUSSION

We analysed the impact of delays in data availability on NRTVSS implementation using CPRD as a way to inform data providers about measures that could improve performance of a NRTVSS system. Our results showed that delays affect power, but only slightly. There were almost no differences observed in the expected time to signal, even when there were improvements in power. Removing delays would thus be insufficient to improve the performance of a system using CPRD data, as the main limiting factor is the volume of data available.

The small differences between each scenario are probably related to the performance measures being calculated on the basis of expected events at the end of the surveillance period. Most individuals are vaccinated at the beginning of the season, and by its end, data have had enough time to accrue. This applies to both sources of delays.

Assessment of the performance at mid-season revealed that we would be able to detect only very large increases in risk at the beginning of December. This raises the issue of timeliness, as by then most individuals would have been vaccinated and any intervention might have limited reach.

Clinical Practice Research Datalink currently collects data from practices using VISION software, but it is expanding to include practices using EMIS software.[8] Presently, there are data from 4.4 million active patients. Initial analysis of EMIS practices indicates an additional 2.6 million active patients (Rachel Williams, personal communication). Assuming this would translate to a similar number of expected events, the new data would amount to approximately 3 expected events, which would be sufficient to detect increases of threefold or more in the risk of GBS following seasonal influenza vaccination. This might not be enough to detect small increases in risk, particularly for rare events. Furthermore, including data from practices using a different software may pose new challenges. For example, the adjustment for delays we proposed is based on the delay distribution observed using data from VISION practices, and it might not be applicable for EMIS practices.[3] Including EMIS practices in a NRTVSS will thus require additional exploration of these data.

In our work, we considered a power of ≥80% as a satisfactory performance. However, GBS can be a severe condition, and when implementing a system, it may be necessary to require higher power level to detect more serious conditions (such as 90%). For existing CPRD data, requirement of 90% power would mean that we could only accurately identify increases in risk ≥5.

**TABLE 2** Expected number of events, power, and expected time to signal under different combination of delays

| Minimum events | RR | Delay Scenario | | | |
|---|---|---|---|---|---|
| | | Reference | Scenario 1 | Scenario 2 | Scenario 3 |
| Season 2013-2014 | | | | | |
| Expected number of events | | | | | |
| — | — | 1.89 | 2.09 | 1.94 | 0.62 |
| Power (expected time to signal in terms of data release) | | | | | |
| 1 | **1.5** | 13 | 13 | 13 | 10 |
| | **2** | 25 | 26 | 25 | 16 |
| | **2.5** | 40 | 42 | 40 | 22 |
| | **3** | 55 (J) | 58 (J) | 55 (J) | 30 |
| | **4** | 78 (J) | 81 (J) | 79 (J) | 44 |
| | **5** | 91 (D) | 93 (J) | 92 (D) | 58 (D) |
| | **6** | 97 (D) | 98 (D) | 97 (D) | 69 (D) |
| | **8** | 100 (D) | 100 (D) | 100 (D) | 85 (D) |
| | **10** | 100 (D) | 100 (D) | 100 (D) | 93 (D) |
| 2 | **1.5** | 14 | 15 | 15 | 11 |
| | **2** | 28 | 30 | 29 | 18 |
| | **2.5** | 44 | 46 | 45 | 27 |
| | **3** | 60 (J) | 62 (J) | 61 (J) | 35 |
| | **4** | 82 (J) | 84 (J) | 83 (J) | 52 (D) |
| | **5** | 93 (D) | 95 (D) | 94 (D) | 65 (D) |
| | **6** | 98 (D) | 98 (D) | 98 (D) | 76 (D) |
| | **8** | 100 (D) | 100 (D) | 100 (D) | 89 (D) |
| | **10** | 100 (D) | 100 (D) | 100 (D) | 96 (D) |
| 4 | **1.5** | 16 | 17 | 16 | |
| | **2** | 33 | 34 | 33 | |
| | **2.5** | 50 (J) | 52 (J) | 50 (J) | |
| | **3** | 65 (J) | 68 (J) | 66 (J) | |
| | **4** | 86 (J) | 88 (J) | 86 (J) | [a] |
| | **5** | 95 (J) | 96 (J) | 95 (J) | |
| | **6** | 98 (J) | 99 (J) | 99 (J) | |
| | **8** | 100 (D) | 100 (D) | 100 (D) | |
| | **10** | 100 (D) | 100 (D) | 100 (D) | |
| Season 2014-2015 | | | | | |
| Expected number of events | | | | | |
| — | — | 1.66 | 1.84 | 1.69 | 0.38 |
| Power (expected time to signal in terms of data release) | | | | | |
| 1 | **1.5** | 12 | 13 | 12 | 9 |
| | **2** | 23 | 25 | 24 | 13 |
| | **2.5** | 37 | 40 | 37 | 18 |
| | **3** | 51 (J) | 55 (J) | 52 (J) | 23 |
| | **4** | 74 (J) | 78 (J) | 75 (J) | 34 |
| | **5** | 88 (J) | 91 (J) | 89 (J) | 44 |
| | **6** | 95 (J) | 97 (J) | 96 (J) | 54 (D) |
| | **8** | 99 (D) | 100 (D) | 100 (D) | 70 (D) |
| | **10** | 100 (D) | 100 (D) | 100 (D) | 81 (D) |
| 2 | **1.5** | 14 | 14 | 14 | 10 |
| | **2** | 26 | 28 | 26 | 16 |
| | **2.5** | 41 | 43 | 41 | 22 |
| | **3** | 55 (J) | 59 (J) | 56 (J) | 29 |
| | **4** | 77 (J) | 81 (J) | 78 (J) | 42 |
| | **5** | 90 (J) | 93 (J) | 91 (J) | 53 (D) |
| | **6** | 96 (J) | 98 (J) | 96 (J) | 63 (D) |
| | **8** | 100 (D) | 100 (D) | 100 (D) | 78 (D) |
| | **10** | 100 (D) | 100 (D) | 100 (D) | 87 (D) |
| 4 | **1.5** | 16 | 16 | 16 | |
| | **2** | 31 | 33 | 31 | |
| | **2.5** | 47 | 50 (J) | 48 | |
| | **3** | 62 (J) | 65 (J) | 63 (J) | |
| | **4** | 83 (J) | 86 (J) | 84 (J) | [a] |
| | **5** | 93 (J) | 95 (J) | 94 (J) | |
| | **6** | 98 (J) | 98 (J) | 98 (J) | |
| | **8** | 100 (J) | 100 (J) | 100 (J) | |
| | **10** | 100 (J) | 100 (J) | 100 (J) | |

Abbreviations: D, December; J, January; RR, rate ratio.

[a]Number of expected events is too small to calculate performance measures.

Our work is subject to some limitations. Our adjustment for recording delays was based on a simplification of the data accrual process and on a historical distribution of delays. Nevertheless, previous work has shown constant recording delay patterns during a 10-year period, which is reassuring.[3] Furthermore, while absence of delays in recording and receiving data is the ideal scenario, it is unlikely that delays in recording can be changed as result of direct action by data providers. Finally, this work is based on a single vaccine/outcome pair. Nevertheless, results for other vaccine/outcome pairs are likely to be similar to the ones observed for seasonal influenza/GBS. The reason for this is twofold: first and as explained above, the lack of improvement in the system's performance is probably related to the fact that the performance is assessed at the end of the surveillance period (when most of the data have already accrued); second, delays in receiving data are fixed and similar for all outcomes. Regarding delays in recording outcomes, GBS is likely to have longer delays than other conditions due to prolonged hospitalisation. Therefore, removing delays in recording these other outcomes would result in even less improvement on power.

In conclusion, minimising delays in data availability are unlikely to substantially improve the performance of a system using CPRD data. Expansion of the data is required.

## ETHICS STATEMENT

All data were anonymised prior to receipt by the authors. Approval for the study was obtained from the Independent Scientific Advisory Committee of the Medicines and Healthcare Products Regulatory Agency (ISAC number: 15_230) and from the Ethics Committee of the London School of Hygiene and Tropical Medicine (LSHTM reference: 10421). The protocol for the overall programme of work was made available for reviewers.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr Ivair Silva, for promptly replying to queries related with the use of the R package Sequential. The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with Public Health England. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health, or Public Health England. The funders had no role in the study design, data collection, analysis, or interpretation.

## CONFLICT OF INTEREST

The authors state no conflict of interest.

## ORCID

*Andreia Leite* http://orcid.org/0000-0003-0843-0630

## REFERENCES

1. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*. 2011;30(1):58-78. https://doi.org/10.1080/07474946.2011.539924
2. Leite A, Andrews NJ, Thomas SL. Implementing near real-time vaccine safety surveillance using the Clinical Practice Research Datalink (CPRD). *Vaccine*. 2017. https://doi.org/10.1016/j.vaccine.2017.09.022
3. Leite A, Andrews NJ, Thomas SL. Assessing recording delays in general practice records to inform near real-time vaccine safety surveillance using the Clinical Practice Research Datalink (CPRD). *Pharmacoepidemiol Drug Saf*. 2017;26(4):437-445. https://doi.org/10.1002/pds.4173
4. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. https://doi.org/10.1093/ije/dyv098
5. Leite A, Andrews NJ, 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(3):225-237. https://doi.org/10.1002/pds.3966
6. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the general practice research database. *Pharmacoepidemiol Drug Saf*. 2005;14(7):443-451. https://doi.org/10.1002/pds.1115
7. Silva IR, Kulldorff M. Sequential: exact sequential analysis for Poisson and binomial data R package version 2.3.1. 2017.
8. The Clinical Practice Research Datalink. Descriptive characteristics of the Clinical Practice Research Datalink (CPRD) EMIS database. 2017; https://www.cprd.com/isac/Protocol_15_217.asp (accessed 13 June 2017).

**How to cite this article:** Leite A, Thomas SL, Andrews NJ. Do delays in data availability limit the implementation of near real-time vaccine safety surveillance using the Clinical Practice Research Datalink? *Pharmacoepidemiol Drug Saf*. 2018;27:25–29. https://doi.org/10.1002/pds.4356