DEPARTMENT OF POPULATION MEDICINE

HARVARD MEDICAL SCHOOL | Harvard Pilgrim Health Care Institute

TREE-BASED DATA MINING FOR COVID-19 VACCINE SAFETY ASSESSMENT IN THE VACCINE SAFETY DATALINK (VSD)

VSD STUDY # 1350

PROTOCOL VERSION: 1.8

PROTOCOL DATE: 01/07/2022

LEAD SITE: Harvard Pilgrim Health Care Institute (HAR)

Principal investigator:	Katherine Yih
Systems engineer:	Judith Maro
Programmer-analysts:	Inna Dashevsky, Robert Rosofsky, David Cole
Project manager:	Jessica LeBlanc
Statistical adviser:	Martin Kulldorff

PROTOCOL CHANGE HISTORY

Version	Date	Change
1.0	02/23/2021	Original protocol
1.1	03/16/2021	Added mention of Janssen vaccine Replaced "approved" with "authorized"
1.2	03/23/2021	Changed follow-up period Made a few other, more minor changes
1.3	04/01/2021	Added detail about signal investigation, including content of line-list of cases to be produced if statistical signals are detected Added title for Table 3 (now Table 2)
1.4	5/26/2021	Added additional details about signal investigation for vaccinees under the age of 20.
1.5	6/04/2021	Additional details about the data we may receive in the instance of signal investigation.
1.6	6/14/2021	Clarifying that we are looking for age group **at vaccination** for signal investigation
1.7	12/02/2021	Removed obsolete analysis plan, including Table 1 Added mention of booster dose analyses Inserted new Level 2 into hierarchical tree of diagnoses Mentioned that cases in outpatient setting would be included in secondary analysis Clarified age groupings for those 90 and older for signal investigation
1.8	01/07/2022	Removed the words "multiple testing" from Background and Methods, given that technically only one hypothesis is being tested Added O codes to list of codes omitted from ICD-10 code tree, with rationale Extended post-vaccination period in which to look at diagnoses and procedures for signal follow-up from 56 days to 70 days (70 days is the maximum follow-up period for the statistical analysis; it was an oversight that a shorter period was specified for signal follow-up in earlier versions of the protocol)

1. BACKGROUND

As of February 2021, the U.S. Food and Drug Administration (FDA) has issued Emergency Use Authorizations for three COVID-19 vaccines, and more are in the pipeline. Post-authorization COVID-19 vaccine safety monitoring is in place, using the Vaccine Adverse Event Reporting System (VAERS), rapid cycle analysis (RCA) of 21 health outcomes by the Vaccine Safety Datalink (VSD), and other data sources and approaches.

We at the Harvard Pilgrim Health Care Institute (HAR) site propose to use a complementary data-mining method, the self-controlled tree-temporal scan statistic, and accompanying software [1] to evaluate whether any of thousands of health outcomes is associated with receipt of COVID-19 vaccines in the VSD population. This method, which builds on earlier work with tree-based scan statistics [2-4], differs from traditional safety study methods in that it does not require pre-specifying a specific health outcome of interest or a specific post-exposure period of potentially increased risk. Instead, for an exposed population, data on all diagnoses recorded within a defined post-exposure follow-up period are scanned to detect any statistically unusual clustering of cases within a large hierarchy, or "tree," of diagnoses as well as within the follow-up period. Evaluation of the thousands of "branches" (e.g., non-traumatic joint disorders) and typically hundreds of time intervals (e.g., Days 19-27 after exposure) is adjusted for by Monte Carlo simulation, conditioning on the number of observations for each diagnosis. Further, the method is self-controlled, eliminating confounding by time-invariant patient characteristics such as chronic disease status.

The method identified known vaccine-associated adverse events and produced few false signals when applied to two vaccines recommended for adolescents and young adults (Menactra [5] and Gardasil [6]) and a vaccine recommended for older adults (Zostavax [7]).

2. STATISTICAL ANALYSES AND PARAMETERS

a. Overview of analyses

We will conduct analyses of the primary COVID-19 vaccination series as well as of booster doses.

b. Study population and enrollment criteria

The study population will consist of VSD enrollees who, at the time of vaccination, are within the age range for which the vaccine is authorized and who are enrolled from at least 400 days prior to vaccination through the respective follow-up period after vaccination (see below).

c. COVID-19 vaccines, dose regimens, and follow-up periods

Primary series: We plan to conduct separate analyses for each FDA-authorized COVID-19 vaccine used by the VSD population, up to five vaccines.

The primary series of some COVID-19 vaccines consists of two doses. The recommended dose-spacing is 21 days for the Pfizer vaccine and 28 days for the Moderna vaccine. For two-dose COVID-19 vaccines, we propose a follow-up period of 10 weeks (70 days) after the first dose, which will capture up to 6-7 weeks after the second dose in most instances. If at a certain point it is decided that Dose 2-specific analyses are needed, those will be added.

For one-dose COVID-19 vaccines, the follow-up period will be 8 weeks (56 days).

Booster doses: Analyses of booster doses will be conducted separately from analyses of the primary series. The follow-up period will be 8 weeks (56 days).

d. Hierarchical diagnosis tree

We will identify outcomes using International Classification of Diseases, Tenth Revision (ICD-10-CM) codes. ICD-10-CM codes have a hierarchical tree-like structure, starting with 21 broad categories of diagnoses, e.g., diseases of the circulatory system, which progressively branch into more and more specific sets of diagnoses, culminating in a highly specific diagnosis code. The ICD-10-CM tree we will use has seven levels. Table 1 presents an example of the hierarchical classification scheme; this example does not use the seventh level.

Table1. Example of hierarchical organization of ICD-10-CM coding system.

Level	Code range or code	Description
1	M00-M99	Diseases of the musculoskeletal system and connective tissue
2	M05-M14	Inflammatory polyarthropathies
3	M06	Other rheumatoid arthritis
4	M06.0	Rheumatoid arthritis without rheumatoid factor
5	M06.01	Rheumatoid arthritis without rheumatoid factor, shoulder
6	M06.011	Rheumatoid arthritis without rheumatoid factor, right shoulder

The primary analyses of the primary series will not use the second level. This level was added to the tree we use for TreeScan analyses recently and will be used in secondary/sensitivity analyses.

We will remove the following groups of diagnoses from consideration. Most are not plausible vaccine-associated adverse events within just a few weeks after vaccination. The O codes would general false signals due to the tendency of pregnant people to get vaccinated later rather than earlier in pregnancy and the self-controlled nature of this analysis method. (Other monitoring efforts, including in VSD, are conducting surveillance for pregnancy-related COVID-19 vaccine adverse events.)

C00-D49 Neoplasms
O00-O9A Pregnancy, childbirth and the puerperium
P00-P96 Certain conditions originating in the perinatal period
Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities
V00-Y99 External causes of morbidity

Z00-Z99 Factors influencing health status and contact with health services

e. Incident diagnoses

For the main analyses, we will define "incident" diagnoses as those observed and recorded in the inpatient or emergency department setting during the post-vaccination follow-up period, as long as the patient was not assigned another ICD-10 diagnosis code having the same first three characters (i.e., in the same second level of tree) in any setting during the prior 400 days. (The choice of 400 days would allow ascertainment of pre-existing conditions recorded at a visit roughly 1 year prior, considering that some patients have annual preventive care visits, although this would have changed to some degree during the pandemic.) So as not to squander statistical power, we do not plan to look for clustering (signals) in the broadest or finest levels of the tree. Incidence will be determined using the second or third level of the tree, above which no analysis of clustering will be carried out, thus no patient will contribute more than one case count to any cluster.

In secondary analyses, we will include cases observed and recorded in the inpatient, emergency department, *or outpatient* setting.

f. Risk and comparison windows

For the primary follow-up period of *x* days, we propose to evaluate case clustering in all intervals between 2 and *x*/2 days long that start on or after 1 day after vaccination (Table2). The comparison period to evaluate each eligible potential risk window will simply consist of the days within the follow-up period that are not in the risk window (Figure 1).

Table2. Parameters for risk windows to be evaluated for two- and one-dose (including boosters) COVID-19 vaccines.

No. of doses in series	**Length of follow-up**	**First day of any cluster evaluated**	**Range of cluster intervals evaluated (in days)**
2 (e.g., Pfizer & Moderna)	70 days after Dose 1	Day 1 after Dose 1	2-35
1 (e.g., Janssen & boosters)	56 days	Day 1	2-28

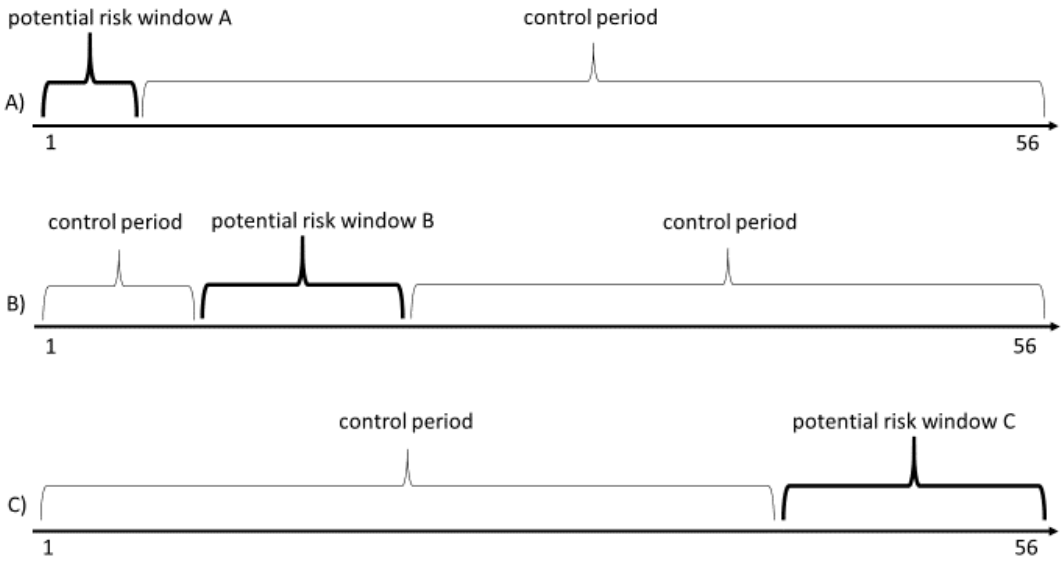

Figure 1. Examples of potential risk windows evaluated at any given instant of the tree-temporal analysis, with their control period(s), assuming a follow-up period of 56 days (to be used for one-dose vaccines). Example A shows a potential risk window that starts on Day 1 after vaccination. The corresponding control period starts the day after the end of the potential risk window and extends through Day 56. Example B shows a potential risk window situated at neither end of the follow-up period but rather somewhere in between. The corresponding control period consists of the segments of the 56-day follow-up period that are not in the potential risk window being evaluated. Example C shows a potential risk window ending on the last day of follow-up. The corresponding control period starts on Day 1 after vaccination and extends through the day before the start of the potential risk window.

g. The tree-temporal scan statistic

With the tree-temporal scan statistic [1], one performs multiple temporal scan statistics, one for each of the many clinical outcomes and groups of related clinical outcomes (i.e., leaves and branches of the tree). For each leaf and branch, one evaluates multiple potential risk windows, comparing the number of events within the risk window with what would be expected by chance if they were randomly and uniformly distributed over time. Under the null hypothesis, there is no unusual temporal clustering of events on any leaf or branch. Under the alternative hypothesis, there is at least one leaf or branch of the tree for which there is a temporal cluster of events during some time interval.

In using the tree-temporal tree-based scan statistic with a self-controlled design, the comparison is within person among time periods. (The rate of any event in unvaccinated people is not measured and is not used for comparison or to standardize any other rates.) The question being asked is whether there is an elevated occurrence of cases of a particular kind of adverse event during a particular time period post-exposure as compared with the rest of the period observed. Rather than pre-specifying the time period of interest for a potential elevation in risk, we allow the data to tell us whether any such period exists. The formula for excess cases is as follows:

(Actual Cases Observed in the Risk Window) – [(Length of the Risk Window)*(Number of Cases Observed Outside the Risk Window / Length of Time Outside the Risk Window)]

The second term (in square brackets) represents the number of cases that would occur in the risk window being evaluated if the cases in the risk window were occurring at the same rate as in the comparison period. The excess is what is observed beyond this value. To calculate attributable risks for statistical signals deemed true indications of vaccine adverse events, we will divide this number by the total number of doses administered.

Monte Carlo simulation will be used to adjust for the evaluation of the thousands of different diagnoses and the hundreds of different time intervals. The number of Monte Carlo replications will be 9999, meaning the lowest possible p-value will be 0.0001. The cut-off p-value to determine statistical significance will be 0.01.

3. SIGNAL FOLLOW-UP

If one or more statistical signals are detected in analysis, we will distribute additional programs to the sites to create and save a snapshot of data for only the patients associated with the signal(s). Only the SAS log and signature data set will be returned to HAR for review to confirm that the program ran without error. Once the data are frozen in this way, there will be time to deliberate about what, if any, follow-up investigation is necessary. Which signals to investigate will be decided collaboratively with CDC and VSD co-investigators.

To investigate a signal, we will provide the sites with programs to generate, for each case contributing to the signal, a list of diagnoses and procedures from the 56 days before through the 56 days after vaccination from the stored patient-level information [8]. VSD co-investigators may choose to review the deidentified patient-level reports solely at their site or, alternatively, together with HAR, CDC, and other VSD co-investigators, using online conferencing with screen-sharing. The comprehensive list is sometimes enough to determine whether or not a case is worthy of concern. For example, there may be related diagnoses/procedures from prior to vaccination suggesting that the condition contributing to the signal pre-dated vaccination, or there may be a recent trauma or surgery that seems likely to explain the condition.

Chart review might be needed to evaluate some statistical signals. Decisions about which signals warrant chart review will be made in consultation with CDC and VSD co-investigators, based on the seriousness of the adverse event and other criteria. If chart review is considered necessary to investigate a signal, HAR will draw on the experience and clinical expertise at the VSD sites to draft and finalize chart abstraction forms and will coordinate collection and analysis of chart data.

Whether signal follow-up involves solely generating and examining the list of diagnoses and procedures or rather also involves chart review, some details of the individual cases will be reported in order to demonstrate whether or not the signal is plausibly related to vaccination. These details will include (i) whether the respective case was confirmed, (ii) evidence, if any, that the case pre-dated COVID-19 vaccination, (iii) evidence, if any, of other plausible causes of the adverse event, and (iv) other diagnoses/symptoms/procedures/medications that help describe

signals, especially those featuring non-specific diagnosis categories, e.g., "other complications of surgical and medical care, not elsewhere classified."

Thus, for each statistical signal, a line-list of cases may be generated and reported containing the following column headings:

a) Case no. (1, 2, 3, 4, etc.)
b) Age group at vaccination (at least 5-years wide for those 20 years and older, with those 90 years and older collapsed into one category; 1-year intervals for those under the age of 20)
c) No. of days to event within follow-up period (1-70 maximum) and/or dates of Dose 1 vaccination and of event
d) In temporal cluster (signal) (as opposed to elsewhere in follow-up period) (Y/N)
e) Day Dose 2 received (where Day 0 is day of Dose 1 receipt) and/or date Dose 2 received
f) Diagnosis code putting case in signal (index diagnosis)
g) Other diagnoses/symptoms/procedures/medications recorded on same day as index diagnosis
h) Diagnoses/symptoms/procedures/medications from prior to vaccination that suggest condition predated vaccination, and number of times each was recorded prior to vaccination
i) Diagnoses/symptoms/procedures/medications from same day as or prior to index diagnosis that suggest condition might have cause other than vaccination
j) Amount of enrolled time prior to vaccination (≥ 400 days by design) and/or enrollment date
k) Confirmed incident case (Y/N)

We will seek clinical expertise from the sites to help quickly generate general lists of diagnoses/symptoms/procedures/medications for (h) and (i) above after detection of the respective statistical signal but *prior* to any examination of cases' records.

4. DATA MANAGEMENT AND QUALITY CONTROL

HAR programmers will create and distribute a standardized set of SAS programs to participating sites to extract data on COVID-19 vaccinees from the VSD data model, convert the data into the appropriate format, and create aggregate datasets for analysis from the converted data. Participating sites will execute these programs and return summary level data to HAR, which will perform TreeScan analytics. HAR has a suite of TreeScan programs ready to be used in this study.

Reiterating from the signal follow-up section above, if a statistical signal is detected, HAR will distribute SAS programs to the sites to save copies of the data on patients contributing to the signal and, if signal investigation is needed, will also send Patient Episode Profile Retrieval (PEPR) programs to these sites. The PEPR reports, consisting of deidentified patient-level data, will be reviewed by participating site PIs, and/or collectively using screen-sharing, to prevent transfer of patient-level data to HAR.

All SAS programs are parameterized, standard programs that have been quality checked before routine use. Additionally, HAR tests all programs on local databases that are appropriately formatted prior to distribution.

5. STRENGTHS AND LIMITATIONS

Strengths of this proposal and approach include the following:

1. The tree-temporal scan statistic does not require pre-specifying a specific health outcome of interest *or* a specific post-exposure period of potentially increased risk, allowing a broad safety assessment.

2. The method uses only data for people with a record of having been vaccinated, eliminating any bias from misclassifying vaccinated people as unvaccinated.

3. We have established procedures for conducting initial signal investigation using electronic data, which may obviate the need for labor-intensive medical chart review in some instances.

4. The proposal is flexible and includes options to separate or combine similar vaccines in analyses, separate or combine Doses 1 and 2 in analyses, consider alternative follow-up periods, select more one than one set of risk window parameters, etc.

5. The analyses will be self-controlled, controlling very well for time-invariant characteristics, including chronic disease status.

6. We will take multiple looks at the data, responding to the need for early safety information.

Limitations include the following:

1. If true adverse reactions do not show strong clustering in time (because of insidious onset) or in the diagnosis tree (because they manifest across multiple body systems and might be coded differently, depending on the case and/or the clinician), they might not be detected. We believe that this would be a concern for only a small minority of potential adverse reactions, however.

2. Time-varying confounding could occur under certain circumstances. For example, if, because of uneven vaccine availability, a large proportion of the study population is vaccinated in May 2021, signals could potentially emerge for adverse events that tend to be more common in the summertime, e.g., sunburn or trauma-related outcomes. However, the maximum follow-up period we are contemplating using is 10 weeks, which would mitigate against time-varying confounding, compared to longer follow-up periods. In addition, we are experienced in spotting instances of time-varying confounding [9].

3. It must be acknowledged that the method is a signal detection (hypothesis-generating) method. False signals may well occur. Thus, when a signal emerges, no conclusion can be drawn about causality without rigorous evaluation of the specific hypothesis that the vaccine is associated with an increased risk of the adverse event identified in the signal.

Essentially, the method serves as a screening tool for identifying possible adverse reactions that must then be investigated further.

6. TIMELINES AND DELIVERABLES

a. Sequence of events

The approximate timing of the main components of the work is shown in Figure 2 below. (We hope to set up the system and start conducting analyses on a slightly shorter timeline than required by the contract, although programming to transform the data into the appropriate format for the existing, quality-checked TreeScan programs to run on will take several months.)

In view of the multiple vaccine products in need of monitoring and the number of different analyses planned, it is possible that difficulties in performing all the analyses and signal follow-up (i.e., bottlenecks) will arise at one or more points during the project period. If this should occur, we will determine the relative priorities of the various analyses and follow-up work in consultation with CDC and VSD co-investigators and carry out the work in order of priority.

Year 1:	1/21	2/21	3/21	4/21	5/21	6/21	7/21	8/21	9/21	10/21	11/21	12/21
Writing	Present study concept		Finish protocol draft	Incorporate input	Finish protocol							
Administrative (monthly meetings, minutes, & status reports ongoing)						Document IRB approval & DUAs						
Data extraction			Write & test SAS programs to convert & extract data			Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data
TreeScan analysis & signal follow-up						First set of analyses of at least Pfizer & Moderna vaccines	More analyses, signal follow-up		Ongoing	Ongoing	Ongoing	Ongoing

Year 2:	1/22	2/22	3/22	4/22	5/22	6/22	7/22	8/22	9/22	10/22	11/22	12/22
Writing												
Administrative (monthly meetings, minutes, & status reports ongoing)												
Data extraction	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data	Extract data
TreeScan analysis & signal follow-up		Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing

Year 3:	1/23	2/23	3/23	4/23	5/23	6/23	7/23	8/23	9/23	10/23	11/23	12/23
Writing								Draft manuscript	Incorporate input	Submit ms. for CDC clearance	Incorporate input	
Administrative (monthly meetings, minutes, & status reports ongoing)												Archive project datasets & provide checklist
Data extraction	Extract data	Extract data	Extract data	Extract data								
TreeScan analysis & signal follow-up		Ongoing	Ongoing	Last analyses	(Continue any necessary signal follow-up)							

Figure 2. Proposed timeline for the epidemiologic project to evaluate the safety of COVID-19 vaccines in the VSD population using tree-based data mining.

b. Administrative commitments

We commit to providing the deliverables in at least as timely a fashion as specified in the RFTOP (Table3):

Table3. Due-dates for deliverables.

Item	Deliverable	Time
1	Provide a concept of the proposed project to CDC for review	Within 60 days of contract award (by 3/8/2021)
2	Provide a draft protocol to CDC for review	Month 4 of contract award (by April 2021)
3	Provide a final protocol to CDC for approval	Month 6 of contract award (by June 2021)
4	Provide documentation to CDC of IRB approval and DUA execution from contributing sites	Month 8 of contract award (by August 2021)
5	Draft manuscript for CDC review	Month 32 of contract (by August 2023)
6	Final manuscript for CDC clearance	Month 34 of contract (by October 2023)
7	Archival of final dataset	Month 36 of contract (by December 2023)
8	Meeting minutes	Within 7 days of meeting
9	Monthly Status Reports	10th of each month

7. REFERENCES

1. Kulldorff M; Information Management Services, Inc. TreeScan: software for the tree-based scan statistic, Version 1.4, with user guide. https://www.treescan.org/. Published 2014. Updated June 2018. Accessed September 28, 2020.
2. Kulldorff M, Dashevsky I, Avery TR, et al. Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol Drug Saf.* 2013;22(5):517-523.
3. Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. *Biometrics.* 2003;59(2):323-331.
4. Brown JS, Petronis KR, Bate A, et al. Drug adverse event detection in health plan data using the Gamma Poisson Shrinker and comparison to the tree-based scan statistic. *Pharmaceutics.* 2013;5(1):179-200.
5. Li R, Weintraub E, McNeil MM, et al. Meningococcal conjugate vaccine safety surveillance in the Vaccine Safety Datalink using a tree-temporal scan data mining method. *Pharmacoepidemiol Drug Saf.* 2018;27(4):391-397.
6. Yih WK,* Maro JC,* Nguyen M, et al. Assessment of quadrivalent human papillomavirus vaccine safety using the self-controlled tree-temporal scan statistic signal-detection method in the Sentinel System. *Am J Epidemiol.* 2018;187(6):1269-1276. *Co-primary.
7. Yih WK, Kulldorff M, Dashevsky I, Maro JC. Using the self-controlled tree-temporal scan statistic to assess the safety of live attenuated herpes zoster vaccine. *Am J Epidemiol*. 2019;188(7):1383–1388.

8. Cole DV, Kulldorff M, Baker M, et al. Infrastructure for evaluation of statistical alerts arising from vaccine safety data-mining activities in Mini-Sentinel. Final report posted to Sentinel website, July 2016. https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_PRISM_Data-Mining-Infrastructure_Report_0.pdf.
9. Yih WK, Kulldorff M, Dashevsky I, Maro JC. A broad safety assessment of the 9-valent human papillomavirus vaccine. *Am J Epidemiol*. In press.