Use of the Tree-Based Scan Statistic for Surveillance of Infant Outcomes Following Maternal Perinatal Medication Use

Sentinel Methods

Elizabeth A Suarez[1], Michael Nguyen[2], Di Zhang[3], Yueqin Zhao[3], Danijela Stojanovic[2], Monica Munoz[4], Jane Liedtka[5], Abby Anderson[8], Wei Liu[7], Steven Bird[7], Inna Dashevsky[1], David Cole[1], Sandra DeLuccia[1], Talia Menzin[1], Jennifer Noble[1], Judith C Maro[1]

1. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 2. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; 3. Office of Biostatistics, Center for Drug Evaluation and Research, FDA, Silver Spring, MD; 4. Division of Pharmacovigilance, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; 5. Division of Pediatric and Maternal Health, Center for Drug and Evaluation Research, US Food and Drug Administration, Silver Spring, MD; 6. Division of Bone, Reproductive, and Urologic Products, Center for Drug and Evaluation Research, US Food and Drug Administration, Silver Spring, MD; 7. Division of Epidemiology, Center for Drug and Evaluation Research, US Food and Drug Administration, Silver Spring, MD; 8. Division of Urology, Obstetrics and Gynecology, Center for Drug and Evaluation Research, US Food and Drug Administration, Silver Spring, MD

Version 2.0

August 28, 2020

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

Use of the Tree-Based Scan Statistic for Surveillance of Infant Outcomes Following Maternal Perinatal Medication Use

Sentinel Methods

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	06/17/2020	Original Version	Sentinel Operations Center
2.0	08/28/2020	• Added a sensitivity analysis to Aim 1 to restrict outcome diagnoses to inpatient setting only, to address outcome misclassification • Extended the study period to March 31, 2019, to include newly available MarketScan data • Added a scenario to the 1:1 matched power calculation simulation analysis in Aim 2 with the sample size of 15000 • Added a bias analysis to the simulation study in Aim 2 to address outcome misclassification • Addition to the Appendices code list: o A list of ICD-10-CM codes included in the infant outcome tree (Appendix Table G) • Addition of appendix in the protocol: o Frequently asked questions and responses about study design decisions and limitations	Elizabeth Suarez

1. Introduction

Pregnant women have historically been excluded from clinical trials during the clinical development of most medical products. As a result, there is often incomplete information about a medical product's safety profile when used during pregnancy. FDA conducts surveillance on the use of medical products in the pregnant population with a specific focus on detecting medical product-induced fetal effects.

Post-marketing requirements have traditionally included establishing a pregnancy registry to monitor drug use (1). Pregnancy registries encounter challenges with recruitment and retention and are often underpowered to find differences in specific malformations. A recent review of registries in the United States reported that the median enrollment was only 36 pregnancies (2). Target sample size is often 300 pregnancies exposed to the drug of interest, however this sample size may only allow for detection of a 2- or 3-fold increase in risk of all of major congenital malformations (MCMs) and is not adequate for detecting an increase in risk in specific malformations (3).

Retrospective, observational studies that utilize electronic health data (EHD, including insurance claims data and electronic health record data) can also be used to evaluate the risk of MCMs and other infant outcomes. However, outcome ascertainment in EHD requires use of a previously validated outcome algorithm in a similar data source, or validation of the algorithm in the intended data source (1). Evaluation of all MCMs as a single outcome may obscure true associations with specific malformations, therefore evaluation of specific outcomes is necessary; this requires validation of many individual outcomes.

Alternatively, the use of signal identification methods in EHD allows for detection of potential increase in risk for all potential MCMs and other important adverse infant outcomes, including preterm birth and low birth weight. Signal identification methods have been used in other areas of pharmacoepidemiology and pharmacovigilance, including monitoring for adverse vaccine effects and for unknown events following initiation of other drugs (4–7). TreeScan™ (http://www.treescan.org) is a statistical data mining tool that can simultaneously scan for increased risk across multiple outcomes and is compatible with multiple study designs (8). It uses a hierarchical outcome tree to group related codes together and applies tree-based scan statistics to adjust for multiple testing when screening across thousands of potential adverse events (8). Use of a hierarchical tree for infant outcomes allows for identification of safety alerts at clinically relevant aggregate groupings (e.g., cardiac malformations) while also testing for potential increased risk of specific outcomes. Observed alerts can then be triaged as known or requiring investigation to determine if the alert was due to bias, confounding, or error (9). Alerts that are potential signals will be evaluated in targeted safety studies specifically designed to quantify the magnitude of effect for a specific health outcome, with confounding control targeted at the outcome of interest, paired with outcome validation, as needed. This approach allows for detection of a wide range of potential adverse effects and focuses rigorous assessments only on alerts that are deemed potential signals.

In this project, we will demonstrate use of a propensity score matched design for TreeScan to identify adverse infant outcomes following maternal exposure to medications during pregnancy.

2. Specific Aims

This is a methods project to evaluate the performance of TreeScan to assess infant outcomes following exposure to medications during pregnancy.

Aim #1: Assess the performance of TreeScan to detect key outcomes in the infant: major congenital malformations, conditions related to gestational duration (e.g., preterm birth), and conditions related to birth weight (e.g., small for gestational age, low birth weight), using empirical data.

Using a propensity score matched design, the TreeScan method will be used to detect potential alerts among mother-infant pairs exposed to fluoroquinolones compared to cephalosporins (referent group) in the first trimester.

Aim #2: Using empirical data to develop background rates, a simulation study will be performed with investigator-injected risks to develop data on the power to detect risk under ideal circumstances.

Using the comparison of first trimester fluoroquinolone or cephalosporin use, background rates of all outcomes in the tree will be estimated. We will assess the power to detect elevated risk under scenarios that vary the sample size per exposure group, the relative risk increase in the fluoroquinolone exposed group, and the baseline prevalence of specified outcomes. Additionally, we will evaluate the impact of fixed 1:N propensity score matching and outcome misclassification on sample size and power. Given that utilization of many medications during pregnancy is rare, the simulation analysis will inform the minimum necessary sample sizes for conducting a TreeScan evaluation and will guide interpretation of results from Aim 1.

3. Case Study: First Trimester Use of Fluoroquinolones and Cephalosporins

As a case study, we evaluate first trimester exposure to fluoroquinolones compared to first trimester exposure to cephalosporins. Potential cases studies were chosen based on the following criteria: 1) older drugs, 2) with well characterized safety profiles for use during pregnancy, and 3) with enough utilization during pregnancy to enable investigation.

Fluoroquinolones are used to treat a variety of infections including urinary tract infections which are common during pregnancy. Quinolones have been shown to be associated with arthropathy in animal models and are contraindicated for use in pediatric and adolescent populations to avoid the risk of musculoskeletal disorders (10,11). Due to these known associations, fluoroquinolones are not widely used during pregnancy (12). While animal models have shown the potential for teratogenic effects (13), results from human studies have not provided strong evidence of an increase in risk for congenital malformations with first trimester fluoroquinolone use. Two meta-analyses reported no association between first trimester quinolone use and birth defects (14,15). Another meta-analysis similarly reported no association between major malformations and quinolones, fluoroquinolones, and ciprofloxacin exposure in the first trimester (16). Results for specific subgroups of major malformations (cardiovascular, genitourinary, nervous system, digestive system) were similarly null (16,17). A recent analysis of US claims data reported that approximately 10% of women with a urinary tract infection in the first trimester were treated with a fluoroquinolone (18).

Cephalosporins are widely used during pregnancy as first-line treatment for multiple infections (13). Studies have shown no association between cephalosporin use and major malformations (19,20), however potential associations with cardiac malformations have been reported by some studies (20–22).

While fluoroquinolones and cephalosporins may be used throughout pregnancy, we are limiting this evaluation to first trimester exposure due to very small sample sizes expected for fluoroquinolone use in the second and third trimesters based on preliminary data on medication utilization by trimester.

4. TreeScan

4.1. Hierarchical Tree for Infant Outcomes

This project will be limited to use of the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coding structure that was initiated in the United States in October 2015. The tree structure is based on the chapters, subchapters, and code structure of ICD-10-CM. Codes from the Q chapter for congenital malformations and the P chapter for conditions originating in the perinatal period were used to define the infant outcomes tree. The leaf level of the tree is comprised of individual ICD-10-CM codes from the Q and P chapters. Individual codes are aggregated into related groups, or nodes, based on the structure of the ICD-10-CM codes, at higher levels of the tree. The ICD-10-CM tree has 6 levels. Nodes at Level 2 of the tree are malformations by body system according to ICD-10-CM subchapters including categories such as “congenital malformations of the circulatory system” and “cleft lip and cleft palate”. Performing hypothesis testing at level 2 mimics groupings of malformations that would commonly be assessed in observational studies using EHD. The tree also allows for hypothesis testing at lower levels that include more specific malformations in each body system. For example, level 3 includes “congenital malformations of cardiac septa” and level 4 includes the code for the critical defect “Tetralogy of Fallot”. Testing at multiple levels of the tree allows for capture of alerts at aggregate groupings while also detecting increased risk of specific malformation types and codes when powered to do so. Using this tree structure also allows for detecting multiple different outcomes that may co-occur if the conditions are not defined by the same incidence criteria. An example of the tree structure is shown in Figure 1.

The tree was further refined to include key outcomes of interest: major congenital malformations, conditions related to gestational duration, and conditions related to birth weight. Codes for minor malformations, genetic conditions, and chromosomal abnormalities were excluded from the tree because they are not outcomes of interest and inclusion may result in major defects in the same node not meeting incidence criteria (see “Defining Congenital Malformation Outcomes” for a description of the incidence criteria). Minor malformation were selected based on guidance from the World Health Organization (WHO)(23). Specific ICD-10-CM codes that could be used to document both major and minor defects were included. The final infant outcome tree contains 6 levels and 1290 leaf level codes. A list of ICD-10 codes included in the outcome tree are included in Appendix Table H.

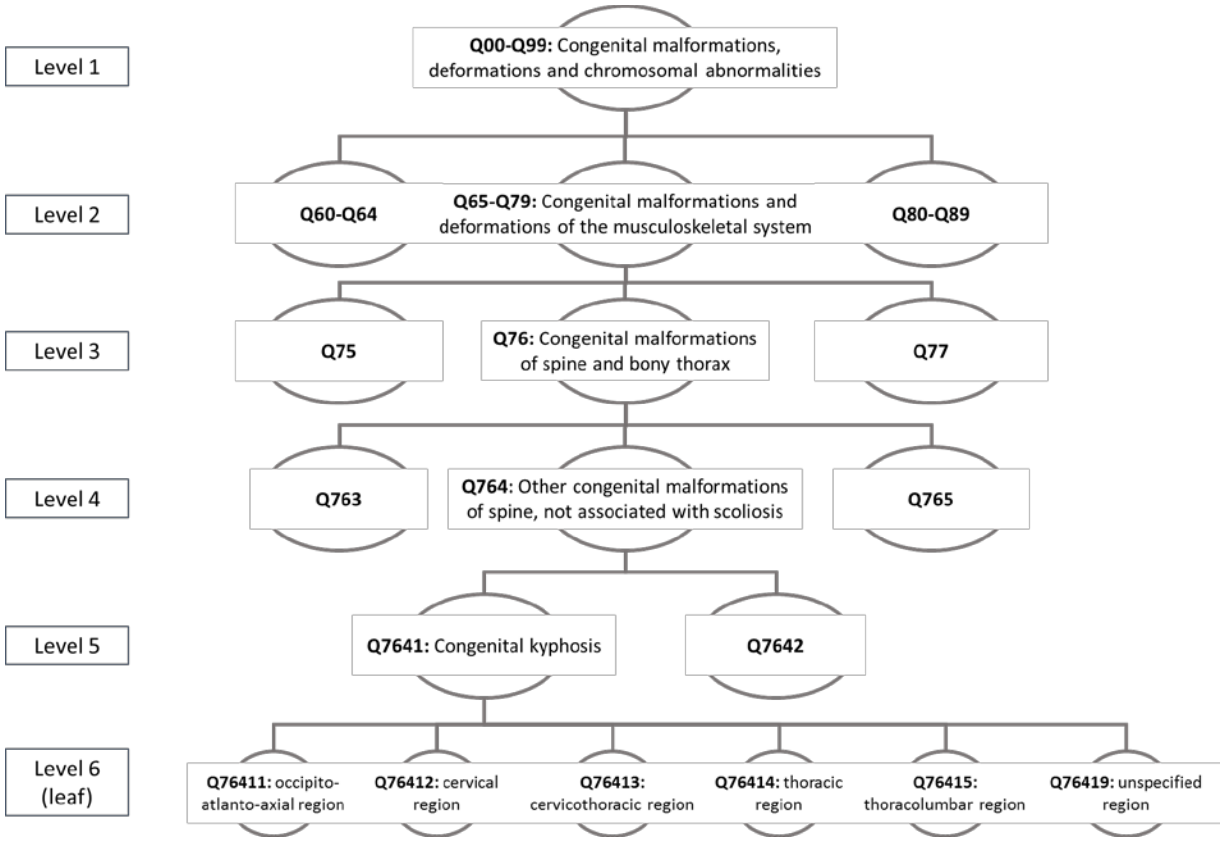

Figure 1. Example from the tailored ICD-10-CM tree for infant outcomes.

4.2. Unconditional Bernoulli Tree Scan Statistic

We will use the unconditional Bernoulli version of the tree-based scan statistic (6). A Monte Carlo based p-value for the test statistic *T* can be obtained by generating random datasets under the null hypothesis that every outcome occurs, independently of other outcomes, with the same probability among in the treatment group versus the comparator group.

The log likelihood ratio (LLR) based test statistic *T* can be calculated as:

$$LLR(G) = ln\left(\frac{\left(\frac{c_G}{c_G + n_G}\right)^{c_G}\left(\frac{n_G}{c_G + n_G}\right)^{n_G}}{(p)^{c_G}(1-p)^{n_G}}\right) I\left(\frac{c_G}{c_G + n_G} > p\right)$$

$$T = \max_G LLR(G)$$

Where: T = unconditional Bernoulli tree scan statistic

c_G = cases in the treatment group for a given node G
n_G = cases in the reference group for a given node G
p = probability of being in the treatment group (for 1:1 matched this is 0.5)
G = node of interest

Random datasets are generated under the null hypothesis by distributing the total number of events per node between the exposed and referent group based on a binomial draw with the expected proportion based on the null hypothesis. When using a 1:1 matched design, this proportion is 0.5. The test statistic

T is calculated for all replicates. The Monte Carlo based p-value is equal to the rank of the test statistic in the real data/(number of replicates+1). If the statistical significance is set to alpha=0.05, then the most likely cut of the real data will be statistically significant if the test statistic ranks in the top 5% of all test statistics from most likely cuts in the real and replicated datasets. This method formally adjusts for multiple hypothesis testing.

5. Aim 1 Methods: Empirical Study

5.1. Data and Study Period

The IBM MarketScan® Research Database will be used for this project. The MarketScan database captures patient-level enrollment, medical, and pharmacy utilization data from predominately large employers and health plans for more than 100 million individuals in the United States. No use of the Sentinel Distributed Database (SDD) is planned for this project. The study period is October 1, 2015 through March 31, 2019; eligible singleton live-birth deliveries that occur during this study period will be included in the analysis. This period was chosen to ensure all deliveries occur in the time period when International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes were used in the United States, enabling use of an ICD-10-CM only outcome tree.

5.2. Creating a Mother-Infant Linkage Table

The Sentinel Common Data Model (SCDM) includes the Mother-Infant Linkage (MIL) table to facilitate the study of infant outcomes following maternal exposures during pregnancy. Eligible live-birth deliveries and infants are linked at each data partner site using available identifiers. The table includes the mother's patient identifier, details on the delivery encounter, and the infant's patient identifier, date of birth, and enrollment information, as well as the method used for linkage (i.e. family subscriber ID, birth certificate, birth registry, etc.). More information on the MIL table can be found on the Sentinel Initiative website (24). This project will be completed using MarketScan data and not using the SDD, therefore an SCDM-compliant MIL table was created for this project.

Live-birth deliveries were identified using ICD-10-CM, ICD-10 Procedure Coding System (ICD-10-PCS), and Current Procedural Terminology, Fourth Edition (CPT-4) diagnosis and procedure codes that indicate live-birth delivery. To follow requirements of the SCDM MIL table, deliveries were eligible for inclusion if they occurred in women aged 10-54 years with a minimum of 180 days of medical coverage prior to the delivery date, and no evidence of a live-birth delivery in the 180 days prior to delivery. Infants were identified by year of birth. We used the linkage criteria utilized by MacDonald, et al. in MarketScan data as a guide for our linkage specifications (25). Live-birth deliveries and infants were linked by family subscriber ID, year of delivery/birth, and when the infant's first encounter date was within 1 day prior to and 30 days after the live-birth delivery date. MarketScan data does not include day and month of birth, therefore the date of birth for the infant was assigned as the live-birth delivery date. Using these criteria, 66% of the live-birth deliveries linked to an infant, similar to the linkage rate reported by MacDonald, et al (25).

5.3. Defining Pregnancy Episodes

For this analysis, we will select singleton live-birth deliveries that are linked to infants from deliveries included in the MIL table. Multiple gestation deliveries will be excluded. To be included in the analysis, linked pairs will be required to have 391 days of maternal medical and drug coverage prior to the date of delivery. This 391-day requirement allows for continuous enrollment during a 90-day pre-pregnancy period and accounts for the longest duration pregnancy episode of 301 days. The start of pregnancy was

designated using the validated Medication Exposure in Pregnancy Risk Evaluation Program (MEPREP) algorithm to estimate pregnancy duration (26). This algorithm was validated using ICD-9-CM codes and was updated to include ICD-10-CM codes, including codes for specific weeks of gestation and codes for preterm and postterm delivery. Gestational duration codes have to occur within 7 days of a delivery date in the inpatient care setting. In the absence of gestational duration codes, pregnancy duration will be set to 273 days. Live-birth deliveries will be excluded from the cohort if there was evidence of a prior delivery during the duration of the pregnancy. The study cohort will be further refined by excluding all mother-infant pairs with first trimester exposure to known teratogens (listed in Appendix Table B). Cohort defining criteria are displayed in the design diagram in Figure 2.

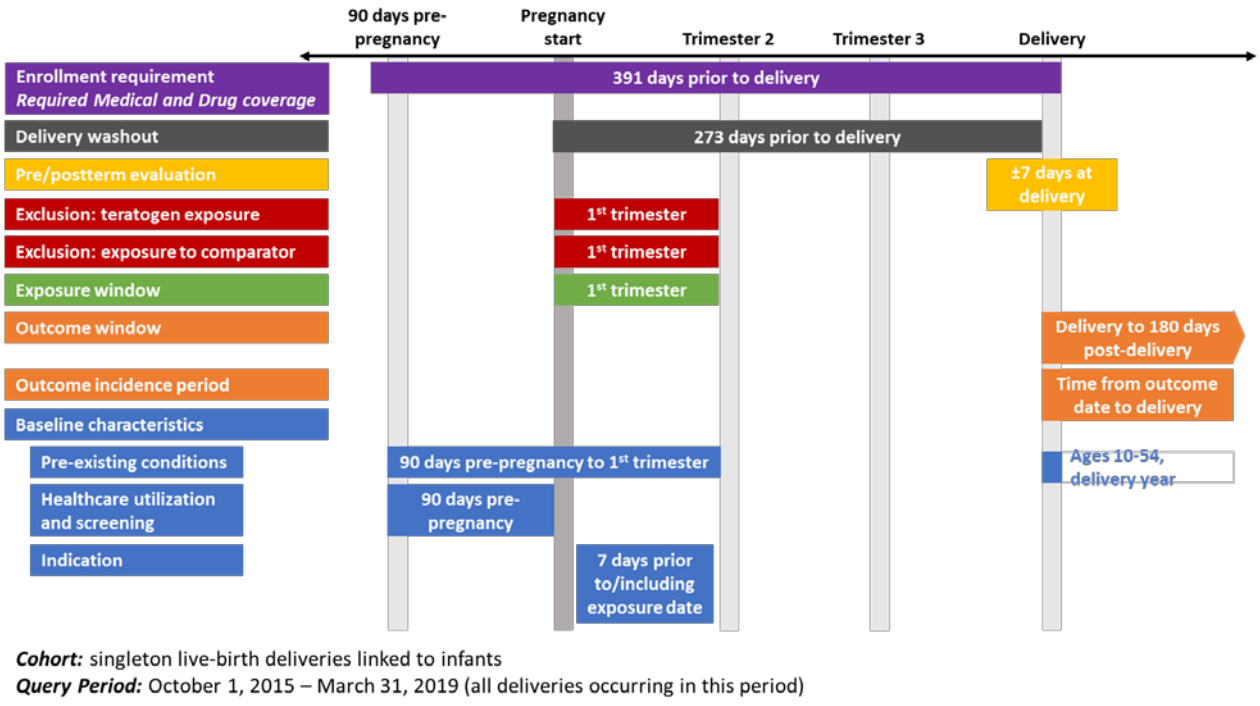

Figure 2. Design diagram for the fluoroquinolone and cephalosporin case study.

5.4. Defining Exposure

We will use Sentinel's routine query tools to extract cohorts with first trimester exposure to fluoroquinolones or cephalosporins in both oral and intravenous forms. National Drug Codes (NDCs) and Healthcare Common Procedure Coding System codes (HCPCS) will be used to define exposure from outpatient dispensing claims and inpatient procedure claims. The fluoroquinolone exposure group will be defined by evidence of prevalent or incident use of a fluoroquinolone in the first trimester without evidence of cephalosporin exposure in the first trimester. The cephalosporin referent group will be defined by evidence of prevalent or incident use of a cephalosporin in the first trimester without evidence of fluoroquinolone exposure in the first trimester. Evidence of exposure will be defined by overlapping days supply; for example, a 7-day prescription that is filled 3 days prior to the start of the first trimester will count as evidence of first trimester exposure because the supply indicates overlap with the start of pregnancy.

5.5. Defining Incident Outcomes

Infant outcomes will be identified using both maternal and infant records. Insurers are required to allow for a special enrollment period of at least 30 days following birth for enrollment of the infant under the parent's insurance (27). Therefore, infants may not have their own patient identification number until days or weeks after birth. Before the infant is enrolled, claims for the infant may appear in the mother's record. To capture all possible outcomes that occur immediately following birth, it is necessary to review both the mother's and infant's records.

Outcomes will be assessed for each mother-infant pair from the delivery date through 180 days after delivery. Outcomes will be included from any care setting.

Outcome incidence will be assessed for each mother-infant pair. The incidence criterion prevents double counting of the same condition in the same mother-infant pair that is evaluated multiple times during the outcome window. The incidence period will be defined as the minimum of the length of the outcome period and the number of days between the outcome date and delivery. This allows for the incidence period to begin at delivery and will not remove outcomes that are diagnosed at delivery but appear in the mother's record prior to delivery as part of prenatal diagnosis and screening.

We will define incident outcomes based on level 3 nodes across the ICD-10-CM tree hierarchy. Incident outcomes will be defined by the first code from the node that occurs on the delivery date or within the outcome window, without any codes in the same level 3 node in the period between the delivery date and the outcome date in any care setting. Multiple incident outcomes may be observed for each mother-infant pair given they meet the incidence criteria at level 3 nodes. Sensitivity analyses will test for alerts at tree level 2, therefore incidence will be established at level 2 for sensitivity analyses.

Mother-infant pairs will be censored at death, disenrollment, or the end of the outcome window. If one member of a 1:N propensity score matched set is censored, the other members will also be censored at the same time.

Given the potential for outcome misclassification when using an inclusive definition for infant outcomes, we will conduct a sensitivity analysis limiting outcomes to the inpatient setting. This analysis will use the same incidence criteria as the main analysis, which requires that the outcome code be the first code from the node that occurs on the delivery date or within the outcome window, without any codes in the same level 3 node in the period between the delivery date and the outcome date in any care setting.

Outcome incidence defined in any care setting and in inpatient only settings can be compared to national reporting on the incidence of specific defects as an informal check for misclassification.

5.6. Propensity Scores

5.6.1. Variables to be included in the propensity scores

The TreeScan method simultaneously tests multiple outcomes, therefore variables for the propensity score cannot be tailored to each exposure-outcome pair. Instead, we established a list of baseline characteristics, pre-existing conditions, screening codes, and healthcare utilization metrics to create a reusable general propensity score that can be used in all propensity score matched TreeScan analyses in pregnancy. The use of a general propensity score for TreeScan analyses in the general population is being assessed in an ongoing Sentinel project (28). We adapted the predefined general score created in that project to be applicable to a pregnant population.

A list of pre-existing conditions was compiled using the pre-existing conditions considered for the Obstetric Comorbidity Score, which predicts severe maternal comorbidity and mortality (29). The list was further refined by adding conditions known to be risk factors for malformations, as suggested by

members of the workgroup. Screening activities were limited to those appropriate for reproductive aged women. A listing of each variable to be included in the general propensity score is in Table 1.

Table 1. Variables to be included in the general propensity score for pregnancy analyses

Category	Source	Variables
Demographics	NA	Age, year of delivery, race and ethnicity[1]
Pre-existing conditions	Bateman (29), workgroup recommendations	Obesity, preexisting hypertension, preexisting diabetes, asthma, drug abuse, alcohol abuse, tobacco use, cardiac valvular disease, chronic congestive heart failure, chronic ischemic heart disease, chronic renal disease, congenital heart disease, cystic fibrosis, HIV, pulmonary hypertension, sickle cell disease/thalassemia, systemic lupus erythematosus, previous cesarean, end stage liver disease, rheumatoid arthritis, inflammatory bowel disease, leukemia/lymphoma, epilepsy/seizure, and psychiatric conditions
Screening	Wang (28,30)	Vaccine administration, Screening examinations and disease management training, Pap smear, HPV DNA test, Fecal occult blood test
Healthcare utilization	Wang (28,30)	Number of inpatient encounters, number of outpatient encounters, number of emergency department visits, number of filled generics

[1]While race and ethnicity are recommended for inclusion in the general propensity score, these variables are not recorded in MarketScan and therefore will not be included in the propensity score for this project.

Prior work on use of a general propensity score versus a tailored score or choosing variables based on an exposure-based high-dimensional approach has demonstrated that the global score is adequate when an appropriate active comparator is used (28). Use of an appropriate active comparator controls for much of the confounding between the exposure and outcome by design. However, it is not always possible to identify a good active comparator when assessing medications used during pregnancy, as women are often channeled into using a drug that is known or suspected to be safe, resulting in little to no use of comparator drugs or use limited to unrepresentative populations (i.e., severe cases). Instead, use of an active comparator with some degree of mismatch on indication or an unexposed referent group will be necessary. To minimize unmeasured confounding, it may be necessary to augment the general propensity score with variables tailored to the drug and referent populations under analysis.

For the case study of fluoroquinolones compared to cephalosporins, we will consider addition of the following variables to the propensity score to define indications for these antibiotics: urinary tract and kidney infections, lower respiratory tract infections, ear, nose, and throat infections, gastrointestinal infections, and sexually transmitted infections. Distribution of these variables in each antibiotic exposure group will be examined prior to addition to the propensity score model.

Additionally, some variables included in the general propensity score should be excluded when sample size is expected to be very small to avoid issues of convergence of the propensity score. The final propensity scores used for this project will be determined using descriptive statistics for the fluoroquinolone exposure group and variables with 0 or very small cells will not be included in the propensity score models.

The evaluation window to be used for each covariate category is illustrated in Figure 2.

5.6.2. Propensity score matching

The propensity score matched cohort design has been used by the FDA Sentinel Program in active surveillance activities and is currently being used for assessment of adverse infant outcomes following maternal exposure to medications during pregnancy in retrospective cohort studies. The use of 1:1 propensity score matching for TreeScan has also been demonstrated in a prior simulation study (7).

We will use 1:1 propensity score matching with various iterations of the propensity score model to control for measured confounding. The matching algorithm will use nearest neighbor matching with a caliper of 0.05.

- Base model: all variables selected for the general propensity score (Table 1)
- Indication model: Base model + the antibiotic indication variables
- High-dimensional propensity score (hdPS) model: variables will be chosen for the propensity score empirically based on their association with the exposure

We will also implement 1:N fixed ratio matching to demonstrate the impact on sample size when requiring >1 match from the referent group. Nearest neighbor matching with a caliper of 0.05 will be used. The number of referent group matches (N) will be dictated by the sample size in the cephalosporin cohort. For example, if the cephalosporin cohort is at least 3 times the size of the fluoroquinolone cohort, we will implement both 1:2 and 1:3 fixed ratio matching.

The distribution of covariates included in the propensity score will be evaluated before and after matching to assess imbalance.

5.7. Identifying Alerts Using TreeScan

In the main analysis, hypothesis testing will be performed at levels 3, 4, and 5. In sensitivity analyses, hypothesis testing will also be performed at level 2. Hypothesis testing will not be done at level 6 (the leaf level) because these codes are primarily used to designate laterality and specific location of a malformation and this level of detail is not informative for identifying specific adverse infant outcomes. The threshold for alerting will be $p \leq 0.05$ (1-sided).

This project is intended to be a methods evaluation rather than a regulatory safety analysis of fluoroquinolone use during pregnancy. Alerts will be triaged as known, expected, or requiring further investigation based on the prescribing information for fluoroquinolone drugs and the known safety profile as documented in the literature.

6. Aim 2 Methods: Simulation Study

6.1. Power using 1:1 propensity score matching

Small sample sizes (<5000 exposed women) are likely to occur when studying medications used during pregnancy. TreeScan may be underpowered to identify signals in these small samples unless the relative increase in risk is very large or the outcome is common. In order to assess the ability of the TreeScan method to detect elevated risk of infant outcomes, we will perform a simulation study with known investigator-injected increases in risk.

Empirical data will be used to inform outcome incidence in our simulated datasets. Outcome counts in the cephalosporin cohort, using all cohort defining criteria used in the empirical study (see the design

diagram in Figure 2), will be used to create the simulated datasets. Exposed and referent cohorts of equal size will be created to mimic a 1:1 propensity score matched scenario.

We will vary the following parameters for each scenario. Sample parameters are noted in Table 2.

- Sample size of the exposed and referent cohorts
- Prevalence of the outcome node with investigator-injected risk
- Magnitude of the relative risk of the investigator-injected risk

For each scenario, we will report significant signals using a threshold for alerting of $p \leq 0.05$ and the power of the dataset to generate an alert.

Table 2. Scenarios to be assessed in the simulation study

Prevalence of outcome	Relative increase in risk in the fluoroquinolone cohort	Sample size of each exposed/referent cohort
Approximately 1 per 10,000	1.5	2000
Approximately 1 per 1,000	2.0	4000
Approximately 1 per 100	4.0	8000
		15000

6.2. Power using fixed 1:N propensity score matching

Using TreeScan in 1:1 propensity score matched populations has been shown to be a valid way to identify signals while controlling for confounding (7). However, use of 1:1 propensity score matching may greatly restrict the sample size available for analysis when the exposed population is small by restricting otherwise large unexposed or comparator exposed referent cohorts. Use of fixed 1:N matching could increase power by increasing the size of the referent cohort as long as the size the exposed cohort does not substantially decrease as patients that have less than N matches are excluded from the cohort. We will evaluate the impact of the use of fixed 1:N matching on sample size and power by simulating commonly observed propensity score distributions and injecting known outcome risks into the resulting matched populations.

Two base scenarios will be selected varying the sample size of the exposed and referent cohorts before matching. We will simulate propensity score distributions in the exposed and referent cohorts with varying levels of overlap. Random samples of the simulated propensity score distributions will be taken to meet the specified unmatched sample sizes, and various fixed matching ratios will be implemented using nearest neighbor matching with a caliper of 0.05. Using the resulting exposed and referent cohort sizes, we will estimate the power to detect a known investigator-injected increase in risk.

6.3. Outcome misclassification

Use of a single diagnosis code to define an outcome is likely to have high sensitivity but may have low specificity and a low positive predictive value (PPV). This type of misclassification is expected to bias relative effect estimates towards the null when it is nondifferential with respect to the exposure, which is a reasonable assumption in an active comparator analysis.

Algorithms in claims data that use multiple codes or concepts to define an outcome are designed to have very high PPV because relative risk estimates will be unbiased when outcome specificity is perfect, even if sensitivity is low, given that any misclassification is nondifferential. However, an algorithm with a high PPV may result in low to moderate sensitivity and lower outcome prevalence because true cases

may be missed by a more restrictive outcome definition (1). This creates a tradeoff between the sensitivity and PPV and the choice to prioritize sensitivity or PPV must consider the study design and objective.

In a TreeScan analysis, a missed signal could be the result of either a) outcome misclassification resulting in bias towards the null, or b) low outcome prevalence resulting in a lack of power to detect an increase in risk. Use of a highly specific outcome algorithm may preserve a true relative increase in risk, but if this algorithm is very restrictive and results in a large drop in prevalence of the outcome, TreeScan may not be powered to detect the alert.

Given the design of TreeScan to evaluate thousands of outcomes simultaneously, it is not feasible to use tailored outcome definitions. However, the tradeoff between sensitivity and PPV can be examined via simulation provided appropriate assumptions are made about sensitivity and PPV. To assess the impact of varying PPV and sensitivity on the ability of TreeScan to detect a true alert, we will preform a simple bias analysis using the simulated data. Assuming the simulated data represents the true data, we will vary PPV and sensitivity to create scenarios with varying levels of misclassification for a single outcome with a known investigator-injected increase in risk. We will report the misclassified incidence and relative risk, and report the power to detect the misclassified relative risk for each scenario. In this exercise, we can evaluate whether misclassification due to imperfect PPV or imperfect sensitivity has a greater impact on our ability to detect true increases in risk, which can inform decisions on the most appropriate outcome definition to use in TreeScan evaluations of adverse infant outcomes.

7. Future Considerations

The current protocol will address first trimester exposure, however future evaluations may also require evaluation of medication exposures in the second and third trimesters. Sample sizes for second and third trimester exposures may be lower than the sample size for first trimester exposures if women discontinue medication use after pregnancy recognition. The power calculations completed in the current protocol will help to inform whether TreeScan is appropriate for second and third trimester exposures.

Additionally, evaluating second and third trimester exposures requires adjustments to the study design to avoid bias that could result in missed signals. Due to birth occurring at different gestational ages, the length of the assessment window for second and third trimester exposures is not uniform across all pregnancies included in a study. Pregnancies with shorter gestations have less opportunity for exposure than pregnancies with longer gestations; this results in exposure appearing to be protective against outcomes associated with shorter gestations (31,32). In single outcome studies, a recommended strategy for avoiding this bias is to use a time-varying exposure definition (32). Use of a time-varying exposure definition is not compatible with the Bernoulli TreeScan statistic, therefore other approaches, such as matching on gestational age of exposure or changing the evaluation window to count back from delivery, could be utilized. The most appropriate way to mitigate this potential for bias when evaluating second and third trimester exposures will be explored in future work.

8. References

1. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Food and Drug Administration, U.S. Department of Health and Human Services. Postapproval Pregnancy Safety Studies, Guidance for Industry, Draft Guidance [Internet]. 2019. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postapproval-pregnancy-safety-studies-guidance-industry

2. Bird ST, Gelperin K, Taylor L, Sahin L, Hammad H, Andrade SE, et al. Enrollment and Retention in 34 United States Pregnancy Registries Contrasted with the Manufacturer's Capture of Spontaneous Reports for Exposed Pregnancies. Drug Saf. 2018 Jan;41(1):87–94.

3. Gelperin K, Hammad H, Leishear K, Bird ST, Taylor L, Hampp C, et al. A systematic review of pregnancy exposure registries: examination of protocol-specified pregnancy outcomes, target sample size, and comparator selection. Pharmacoepidemiol Drug Saf. 2017;26(2):208–14.

4. Yih WK, Maro JC, Nguyen M, Baker MA, Balsbaugh C, Cole DV, 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 Jun 1;187(6):1269–76.

5. 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 Jul 1;188(7):1383–8.

6. Kulldorff M, Dashevsky I, Avery TR, Chan AK, Davis RL, Graham D, et al. Drug safety data mining with a tree-based scan statistic: Tree-Based Scan Statistic Data Mining Method. Pharmacoepidemiol Drug Saf. 2013 May;22(5):517–23.

7. Wang SV, Maro JC, Baro E, Izem R, Dashevsky I, Rogers JR, et al. Data Mining for Adverse Drug Events with a Propensity Score-matched Tree-based Scan Statistic. Epidemiology. 2018;29(6):895–903.

8. Kulldorff M, Fang Z, Walsh SJ. A Tree-Based Scan Statistic for Database Disease Surveillance. Biometrics. 2003 Jun;59(2):323–31.

9. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman JE, Anderson-Smits C, et al. Mini-Sentinel methods: framework for assessment of positive results from signal refinement. Pharmacoepidemiol Drug Saf. 2014;23(1):3–8.

10. Jackson MA, Schutze GE, Committee on Infectious Diseases. The Use of Systemic and Topical Fluoroquinolones. Pediatrics. 2016 Nov;138(5):e20162706.

11. Food and Drug Administration. FDA Drug Safety Communication: FDA updates warnings for oral and injectable fluoroquinolone antibiotics due to disabling side effects [Internet]. 2016. Available from: https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-fda-updates-warnings-oral-and-injectable-fluoroquinolone-antibiotics

12. Food and Drug Administration. Ciprofloxacin Use by Pregnant and Lactating Women [Internet]. 2017. Available from: https://www.fda.gov/drugs/bioterrorism-and-drug-preparedness/ciprofloxacin-use-pregnant-and-lactating-women

13. Bookstaver PB, Bland CM, Griffin B, Stover KR, Eiland LS, McLaughlin M. A Review of Antibiotic Use in Pregnancy. Pharmacother J Hum Pharmacol Drug Ther. 2015;35(11):1052–62.

14. Ziv A, Masarwa R, Perlman A, Ziv D, Matok I. Pregnancy Outcomes Following Exposure to Quinolone Antibiotics – a Systematic-Review and Meta-Analysis. Pharm Res. 2018 May;35(5):109.

15. Yefet E, Schwartz N, Chazan B, Salim R, Romano S, Nachum Z. The safety of quinolones and fluoroquinolones in pregnancy: a meta-analysis. R Coll Obstet Gynaecol. 2018;8.

16. Acar S, Keskin-Arslan E, Erol-Coskun H, Kaya-Temiz T, Kaplan YC. Pregnancy outcomes following quinolone and fluoroquinolone exposure during pregnancy: A systematic review and meta-analysis. Reprod Toxicol. 2019 Apr;85:65–74.

17. American College of Obstetricians and Gynecologists. Sulfonamides, Nitrofurantoin, and Risk of Birth Defects. Committee Opinion No. 717. Obstet Gynecol. 2017;130:e150-2.

18. Ailes EC, Summers AD, Tran EL, Gilboa SM, Arnold KE, Meaney-Delman D, et al. Antibiotics Dispensed to Privately Insured Pregnant Women with Urinary Tract Infections — United States, 2014. Morb Mortal Wkly Rep. 2018 Jan 12;67(1):18–22.

19. Muanda FT, Sheehy O, Bérard A. Use of antibiotics during pregnancy and the risk of major congenital malformations: a population based cohort study. Br J Clin Pharmacol. 83:2557–71.

20. Czeizel AE, Rockenbauer M, Sørensen HT, Olsen J. Use of cephalosporins during pregnancy and in the presence of congenital abnormalities: A population-based, case-control study. Am J Obstet Gynecol. 2001 May;184(6):1289–96.

21. Crider K, Cleves M, Reefhuis J, Berry R, Hobbs CA, Hu D. Antibacterial Medication Use During Pregnancy and Risk of Birth Defects: National Birth Defects Prevention Study. ARCH PEDIATR ADOLESC MED. 2009;163(11):8.

22. Ailes EC, Gilboa SM, Gill SK, Broussard CS, Crider KS, Berry RJ, et al. Association between antibiotic use among pregnant women with urinary tract infections in the first trimester and birth defects, National Birth Defects Prevention Study 1997 to 2011. Birt Defects Res A Clin Mol Teratol. 2016;106(11):940–9.

23. WHO/CDC/ICBDSR. Birth defects surveillance: a manual for programme managers. Geneva: World Health Organization; 2014.

24. Sentinel Operations Center. Mother-Infant Linkage: Frequently Asked Questions & Appendices [Internet]. 2019. Available from: https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/MIL_FAQs&Appendices.pdf

25. MacDonald SC, Cohen JM, Panchaud A, McElrath TF, Huybrechts KF, Hernández-Díaz S. Identifying pregnancies in insurance claims data: Methods and application to retinoid teratogenic surveillance. Pharmacoepidemiol Drug Saf. 2019;28(9):1211–21.

26. Li Q, Andrade SE, Cooper WO, Davis RL, Dublin S, Hammad TA, et al. Validation of an algorithm to estimate gestational age in electronic health plan databases. Pharmacoepidemiol Drug Saf. 2013;22(5):524–32.

27. U.S. Department of Labor. FAQs on HIPAA Portability and Nondiscrimination Requirements for Workers [Internet]. [cited 2020 Apr 8]. Available from: https://www.dol.gov/sites/dolgov/files/EBSA/about-ebsa/our-activities/resource-center/faqs/hipaa-consumer.pdf

28. Wang S, Gagne J, Maro J, Kattinakere S, Stojanovic D, Eworuke E, et al. A General Propensity Score for Signal Detection Using Tree-Based Scan Statistics. Pharmacoepidemiol Drug Saf. 2019;28(S2):29.

29. Bateman BT, Mhyre JM, Hernandez-Diaz S, Huybrechts KF, Fischer MA, Creanga AA, et al. Development of a comorbidity index for use in obstetric patients. Obstet Gynecol. 2013;122(5):957–65.

30. Wang S, Stojanovic D. Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics [Internet]. Available from: https://www.sentinelinitiative.org/sentinel/methods/development-and-evaluation-global-propensity-score-data-mining-tree-based-scan

31. Huybrechts KF, Bateman BT, Hernández-Díaz S. Use of real-world evidence from healthcare utilization data to evaluate drug safety during pregnancy. Pharmacoepidemiol Drug Saf. 2019 May 10;pds.4789.

32. Matok I, Azoulay L, Yin H, Suissa S. Immortal time bias in observational studies of drug effects in pregnancy. Birth Defects Res Clin Mol Teratol [Internet]. 2014;100(9):658-662 doi.10.1002/bdra.23271.

9. Appendix: Frequently Asked Questions

1. **How does this study address the potential for false negative results (i.e., missed signals)?**

The simulation study is designed to address this concern by evaluating the power to detect alerts given various sample sizes, outcome incidence, and magnitudes of relative risk for the outcome between the exposed and comparison groups. Results of the simulation study will inform whether expected sample sizes in future evaluations of medication use in pregnancy are large enough to allow for use of the TreeScan method.

The medications chosen for the empirical study, fluoroquinolones, are not known to be associated with an increased risk for birth defects and no known risks are noted in the labels for these antibiotics. Therefore, we do not have known safety signals to use as a "gold standard" for comparison to the empirical results. However, it is difficult to choose drugs with a known birth defect risk (e.g., a labeled risk) because products with known risks are intentionally avoided during pregnancy and sample sizes of exposed women would be very small. Given the desire to evaluate the performance of TreeScan for future use in the Sentinel System, it was important to select drugs that had sufficient clinical data in ICD-10-CM rather than studying older drugs with clinical data coded in older terminologies.

2. **Why is this study limited to first trimester exposure only?**

The expected sample size for women exposed to fluoroquinolones in the second and third trimesters is less than 500 pregnancies in the data source used for this study. Given the low prevalence of many infant outcomes included in the outcome tree, this sample size is not expected to be adequately powered to detect increases in risk.

Alternatively, we could have defined exposure as any use during pregnancy and avoided issues with small sample sizes in the second and third trimesters. This approach is not recommended because the risk of adverse infant outcomes following medication exposure is not the same throughout the entire gestational period. Inclusion of exposed time that is not at risk for the outcome would attenuate relative effect estimates and could result in missed signals. Therefore, exposure assessment will be limited to first trimester for this analysis.

The feasibility of investigating exposure during any trimester or gestational period should be evaluated prior to starting a signal identification exercise for any medication given the pattern of drug discontinuation after pregnancy recognition; for many medications, exposure prevalence may drop substantially from first trimester to second trimester depending on the indication and utilization patterns of the medication.

3. **Classifying exposure as any exposure in the first trimester could result in attenuated risk for outcomes where the etiologically relevant window is much shorter (e.g., weeks 5-8 for cardiac defects). This could lead to missed signals.**

A signal identification assessment evaluates risk of thousands of outcomes simultaneously, therefore it is not feasible to tailor the exposure window to the most appropriate gestational period of exposure for every outcome. Further, the etiologically relevant window is unknown for many adverse infant outcomes including preterm birth and small for gestational age. Given that first trimester is a critical period of development for many organ systems, this gestational period is commonly evaluated when studying major malformations. When sample size allows, multiple exposure windows can be evaluated. It is also important to consider that the potential for exposure misclassification may increase as the exposure window is shortened due to estimation of gestational age using an algorithm.

However, we could use temporal scans in sensitivity analyses for select outcomes to help identify shorter exposure windows for certain outcomes that alert or are near the alert threshold. Use of scan statistics for this purpose have been demonstrated in a study of vaccine safety (1).

4. **How accurate is the algorithm for classifying gestational age? Inaccurate dating of the pregnancy can result in exposure misclassification.**

Gestational age at delivery is estimated using a validated algorithm that has been adapted to include ICD-10-CM codes for gestational age (2). The validation study for this algorithm assessed the potential for exposure misclassification by comparing classification of first trimester antibiotic use according to the algorithm to a classification according to gestational age from birth certificates. They reported the sensitivity and specificity of first trimester exposure to antibiotics as over 92%. Therefore, we expect misclassification of first trimester exposure due to misspecified gestational age to be minimal.

5. **How will identified alerts and potential false positive results be addressed?**

As stated in the protocol, alerts will be triaged as known, expected, or requiring further investigation based on the prescribing information for fluoroquinolone drugs and the known safety profile as documented in the literature. Alerts requiring further investigation can be evaluated in the following ways. First, the patient episode profile retrieval (PEPR) tool can be used assess the claims profile for a patient to identify sources of confounding. Second, a targeted safety study for a specific outcome could be initiated using a validated algorithm or including a validation study and carefully considered confounding control.

6. **Including only live-birth deliveries in the study population may result in missing outcomes in pregnancies that do not end in live birth. How will that impact results?**

Restricting the study population to live births may result in an undercounting of outcomes, particularly severe birth defects, that are likely to result in pregnancy loss or termination. Pregnancies that do not end in live birth are difficult to accurately identify in administrative data and the reasons for fetal demise or termination are unlikely to be documented. This undercounting of some outcomes will only result in biased relative risk estimates if the exposure is also associated with pregnancy loss or termination. In other words, the estimate would only be biased if the rate of live-birth differs between study groups. Methods are available to quantify the potential impact of missing non-live births and

assess the difference required to dilute a potential meaningful increase identified. In this study, it is unlikely that the risk of pregnancy loss or termination will differ between women using fluoroquinolones or cephalosporins. Therefore, we don't not expect results to be biased due to restriction of the study population to live births only.

7. **Pregnancy and family history may also be associated with an increase in risk for adverse infant outcomes. Should these variables be considered for inclusion in the propensity score?**

While pregnancy and family history of adverse pregnancy outcomes may be predictive of some adverse outcomes in the current pregnancy, we are unable to accurately measure these potential confounders in claims data. Women included in the study population are only required to have 391 days of medical and drug coverage prior to the delivery date to ensure that preexisting conditions can be assessed in the 90 days before the start of pregnancy. Requiring additional coverage prior to the start of pregnancy to allow for capture of previous pregnancy outcomes documented in claims would greatly reduce the available sample size and potentially reduce generalizability of the study population.

References

1. Li L, Xu S, Yan L, Kawai T, Benitez GV, Hua W. Evaluation of Scan Statistics for Assessing Vaccine Safety in Pregnancy [Internet]. 2016. Available from: https://www.sentinelinitiative.org/sites/default/files/vaccines-blood-biologics/assessments/Mini-Sentinel_PRISM_Scan-Statistics-for-Assessing-Vaccine-Safety-in-Pregnancy_Report.pdf

2. Li Q, Andrade SE, Cooper WO, Davis RL, Dublin S, Hammad TA, et al. Validation of an algorithm to estimate gestational age in electronic health plan databases. Pharmacoepidemiol Drug Saf. 2013;22(5):524–32.