Sentinel

Surveillance of Adverse Infant Outcomes Following Maternal Medication Use During Pregnancy Using Tree-Based Scan Statistics

Elizabeth Suarez, PhD

On behalf on the Sentinel TreeScan in Pregnancy Workgroup

elizabeth_suarez@harvardpilgrim.org

Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

Acknowledgements

TreeScan in Pregnancy workgroup members:

Sentinel Operations Center
- Judy Maro
- Elizabeth Suarez
- Sandra DeLuccia
- Jennifer Noble
- Inna Dashevsky
- Talia Menzin
- David Cole

FDA
- Michael Nguyen
- Danijela Stojanovic
- Monica Munoz
- Abby Anderson
- Yueqin Zhao
- Di Zhang
- Jane Liedtka
- Wei Liu

Contents

4 **Section 1**
Motivation

9 **Section 2**
Review of TreeScan Methods

18 **Section 3**
Simulation Analysis

35 **Section 4**
Case Study

51 **Section 5**
Conclusions

Motivation

Why are we interested in using signal identification methods for drug safety in pregnancy research?

Monitoring of pregnancy exposures

- Pregnant women are rarely included in clinical trials during drug development, therefore data on teratogenicity and other potential adverse effects are collected post-market
- Pregnancy Exposure Registries are a primary source of post-market data
  - Pregnancy Exposure Registries often miss enrollment targets
  - Registries are often underpowered for individual malformations (Gelperin, 2018)
- Healthcare utilization data can be used for complementary studies

Gelperin K, Hammad H, Leishear K, 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

Signal identification analyses can supplement current practices for monitoring

- Signal identification = systematic evaluation of potential adverse events related to the use of medical products without prespecifying an outcome of interest
  - Allows for detection of new and unsuspected potential safety concerns
- Signal identification can identify potential adverse events to prioritize for targeted study when there are not known specific safety concerns
- Advantages:
  - Utilize the large sample sizes available in administrative data
  - Not limited to major congenital malformations as a primary outcome – can scan for all types of malformations individually and in clinically relevant groupings (e.g., atrial septal defect, any cardiac malformation)

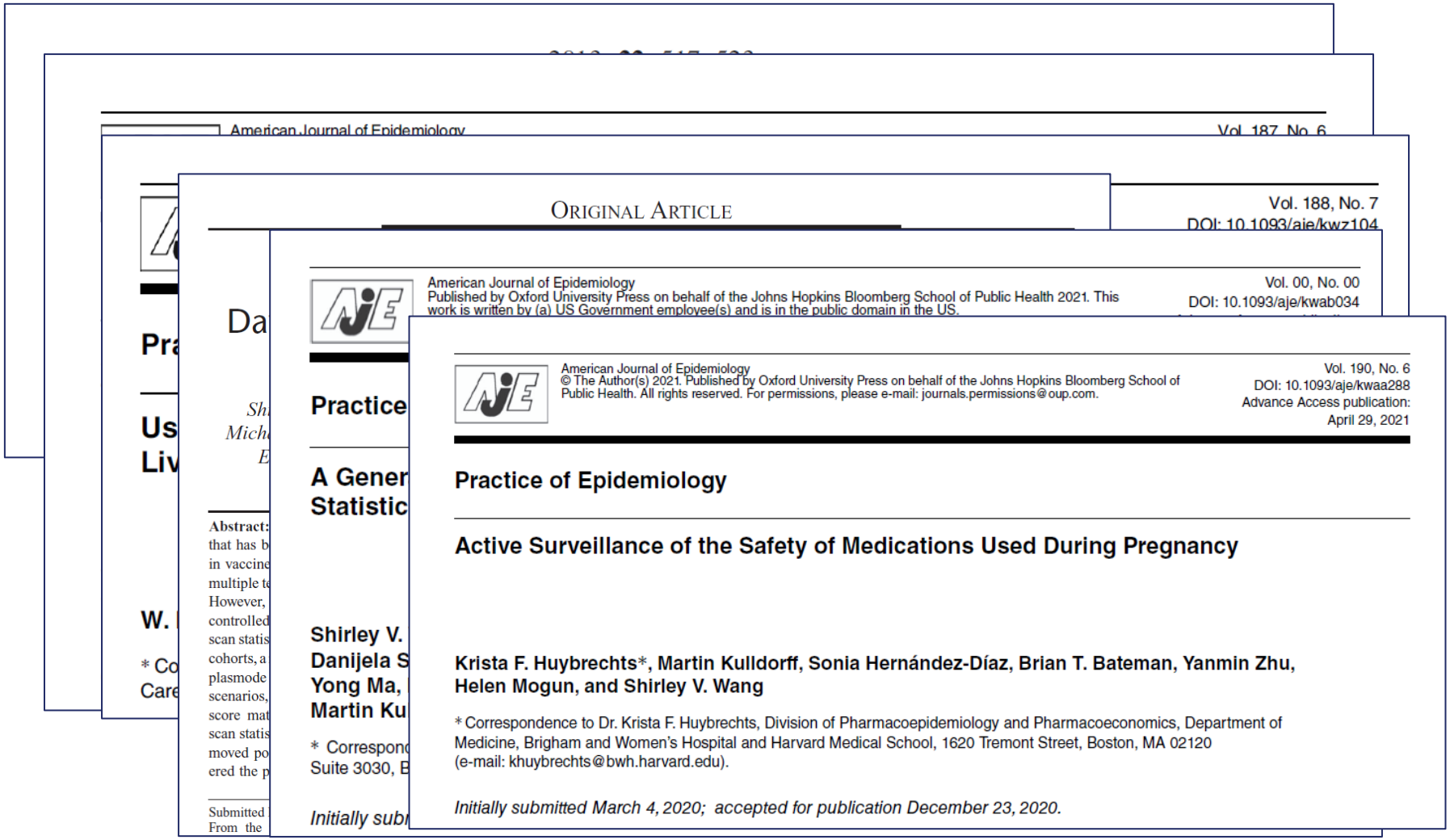
Review of TreeScan methods

Multiple outcome study designs and the TreeScan tool

TreeScan™

- TreeScan is a statistical data mining tool that can be used for signal identification in pharmacovigilance/pharmacoepidemiologic analyses
  - Simultaneously scans for increased risk across multiple outcomes and allows for testing of very specific outcomes (e.g., atrial septal defect) or in groupings of concepts (e.g., congenital malformations of the circulatory system)
  - Formally adjust for multiple scenarios with a composite null hypothesis testing to hold type I error due to chance alone at a user-specified threshold
  - Compatible with multiple epidemiologic study designs and confounding control methods

Previous examples of TreeScan for drug safety



American Journal of Epidemiology
Vol. 187, No. 6

ORIGINAL ARTICLE
Vol. 188, No. 7
DOI: 10.1093/aje/kwz104

American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2021. This work is written by (a) US Government employee(s) and is in the public domain in the US.
Vol. 00, No. 00
DOI: 10.1093/aje/kwab034

American Journal of Epidemiology
© The Author(s) 2021. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Vol. 190, No. 6
DOI: 10.1093/aje/kwaa288
Advance Access publication:
April 29, 2021

Practice of Epidemiology

Active Surveillance of the Safety of Medications Used During Pregnancy

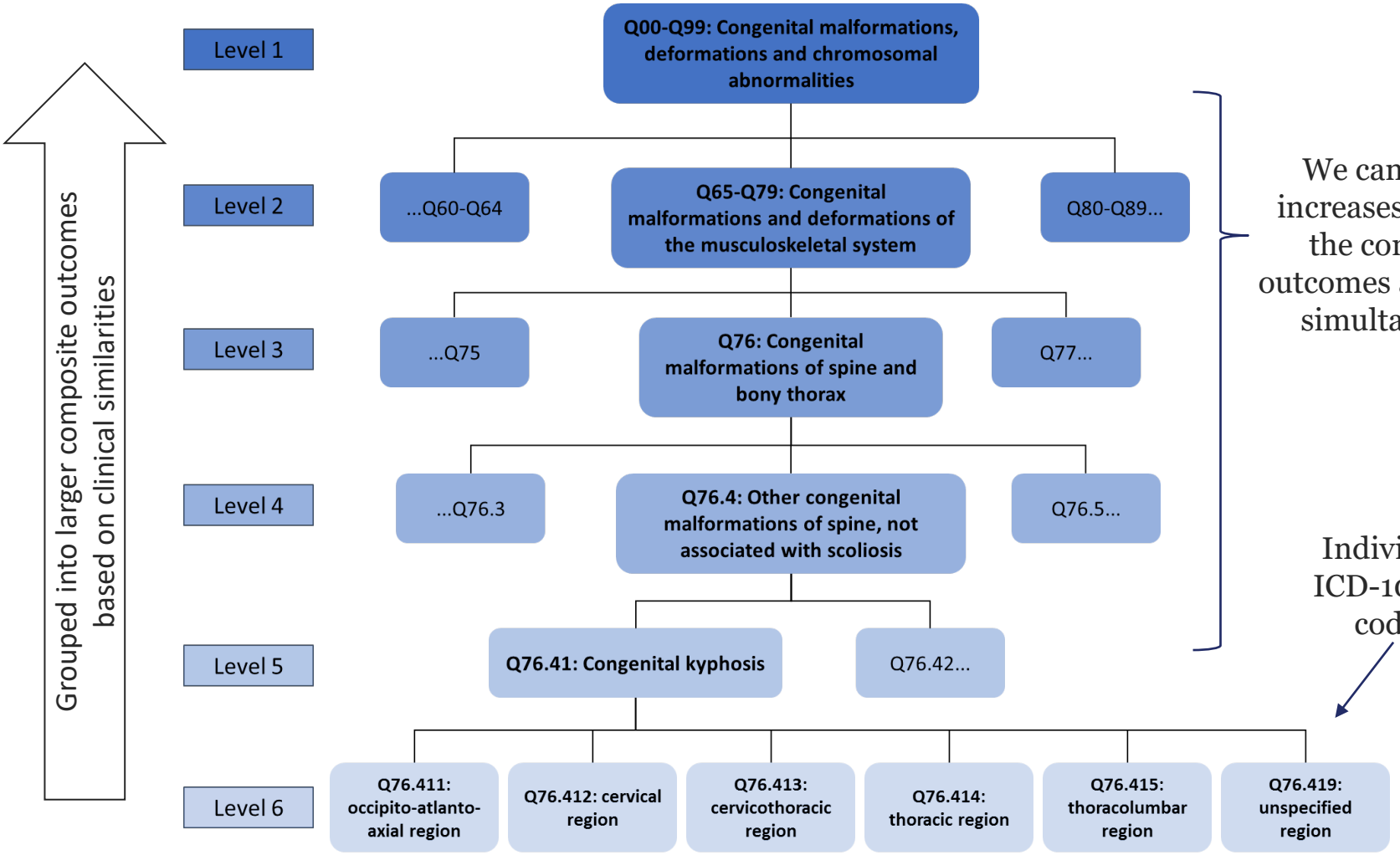
Krista F. Huybrechts*, Martin Kulldorff, Sonia Hernández-Díaz, Brian T. Bateman, Yanmin Zhu, Helen Mogun, and Shirley V. Wang

* Correspondence to Dr. Krista F. Huybrechts, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Boston, MA 02120 (e-mail: khuybrechts@bwh.harvard.edu).

Initially submitted March 4, 2020; accepted for publication December 23, 2020.

How the outcome tree works

In the tree: Major congenital malformations, conditions related to gestational length and birth weight



TreeScan statistics

- The “tree” allows for testing at individual outcomes or related groupings
- The “scan” statistic allows for adjustment of multiple testing across the tree
  - Null hypothesis: there is no increase in risk across any node in the tree
  - Compares observed and expected outcome counts in every node in the tree using log likelihood ratios
  - Selects the maximum LLR as the test statistic and calculates p-values using Monte Carlo hypothesis testing

  ➢This limits the potential for false positive alerts

Study design and confounding control

- Self-controlled designs or cohort designs
  - We're using a cohort design – comparing an exposed to a referent group
  - Need to control for confounding – use propensity score methods
- The LLR can be derived from either a binomial-based or Poisson-based maximum likelihood estimator
  - Bernoulli: assumes all outcomes occur uniformly in a population with a fixed probability of belonging to the exposed group
    - Works well with fixed ratio propensity score matching
  - Poisson: assumes outcomes in the exposed group follow a Poisson distribution based on the outcome rate in the referent group
    - Works well with propensity score stratification

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

Study aims

1. **Simulation study: Assess the performance of TreeScan under known conditions**
   - Can TreeScan identify an increase in risk for a specific malformation in our tree, given a certain sample size?
   - We can simulate a cohort with a known increase in risk to determine if TreeScan is powered to detect pre-specified increases in risk
2. **Case study: Demonstrate the use of TreeScan in real data, in a cohort of pregnant women linked to their live-born infants**
   - How do results look in real data?
   - How do results compare when we use different propensity score methods/TreeScan models?

Simulation analysis

Methods and Results

What can a simulation study teach us?

- **Simulation study: Assess the performance of TreeScan under known conditions**
- We are mainly interested in power: the probability of correctly rejecting the null hypothesis (of no increased risk at any position in the outcome tree)
- Previous simulation studies have estimated power to identify signals with TreeScan, but:
  - were based in the general adolescent and adult populations
  - used a different outcome tree
  - were generally interested in very rare outcomes in large populations
- In pregnancy studies, we often have small exposed populations even in administrative data (e.g., <5000), but composite malformations outcomes are not very rare (approximately 1 per 1000)

What can a simulation study teach us?

- Simulation studies can also be used to answer specific questions about study design options:
  1. Can we increase power by using a different statistical method?
  2. How does our outcome definition impact power, given that outcome misclassification is common in administrative data?

- I'll walk through the methods and results of these questions in this section

General simulation methods

1. Used empirical data to estimate the background incidence of outcomes in our tree
   - IBM MarketScan® Research Database
   - Estimated outcome incidence for each outcome in the tree in an unexposed referent population of pregnant women linked to infants
2. Simulated cohorts with known increases in risk of pre-specified outcomes
   - Selected malformation outcomes with incidence varying from approximately 1 per 10,000 to 1 per 100
   - Increased the risk for that pre-specified outcome by a risk ratio of 1.5, 2, or 4
   - Varied the size of the exposed sample
3. Calculated power to detect the known increase in risk in the simulated cohort using the TreeScan software

Question 1: What is the power to identify signals in scenarios expected in a pregnancy study?

- We are interested in two propensity score methods, which use two different probability models for TreeScan: Bernoulli and Poisson
- Because these models use different methods to calculate the expected number of exposed outcomes and the test statistic, they differ in power
- We estimated power under both models for comparison:
  - Bernoulli
  - Poisson

Power estimates varying TreeScan model, outcome incidence, sample size, and relative risk (RR)

	Bernoulli	RR			Poisson	RR		
	# exposed	1.5	2.0	4.0	# exposed	1.5	2.0	4.0
Incidence = 8 per 1000 Q21.0: ventricular septal defect	2000	0.08	0.25	1.00	2000	0.10	0.50	1.00
	4000	0.11	0.58	1.00	4000	0.21	0.89	1.00
	8000	0.24	0.90	1.00	8000	0.56	1.00	1.00
	15000	0.55	1.00	1.00	15000	0.92	1.00	1.00
	20000	0.75	1.00	1.00	20000	0.98	1.00	1.00
	30000	0.92	1.00	1.00	30000	1.00	1.00	1.00
Incidence = 1.8 per 1000 Q40.0: pyloric stenosis	2000	0.06	0.08	0.37	2000	0.06	0.09	0.72
	4000	0.05	0.08	0.74	4000	0.06	0.16	0.97
	8000	0.06	0.14	0.98	8000	0.10	0.44	1.00
	15000	0.08	0.39	1.00	15000	0.19	0.82	1.00
	20000	0.10	0.56	1.00	20000	0.24	0.93	1.00
	30000	0.15	0.78	1.00	30000	0.44	0.99	1.00
Incidence = 0.6 per 1000 Q35.9: cleft palate, unspecified	2000	0.05	0.05	0.05	2000	0.06	0.06	0.16
	4000	0.05	0.05	0.14	4000	0.05	0.07	0.50
	8000	0.05	0.08	0.34	8000	0.06	0.12	0.85
	15000	0.06	0.09	0.77	15000	0.07	0.23	0.99
	20000	0.06	0.11	0.93	20000	0.08	0.31	1.00
	30000	0.06	0.17	1.00	30000	0.10	0.51	1.00

Power estimates varying TreeScan model, outcome incidence, sample size, and relative risk (RR)

	Bernoulli	RR			Poisson	RR		
	# exposed	1.5	2.0	4.0	# exposed	1.5	2.0	4.0
Incidence = 8 per 1000 Q21.0: ventricular septal defect	2000	0.08	0.25	1.00	2000	0.10	0.50	1.00
	4000	0.11	0.58	1.00	4000	0.21	0.89	1.00
	8000	0.24	0.90	1.00	8000	0.56	1.00	1.00
	15000	0.55	1.00	1.00	15000	0.92	1.00	1.00
	20000	0.75					.00	1.00
	30000	0.92					.00	1.00
Incidence = 1.8 per 1000 Q40.0: pyloric stenosis	2000	0.06					.09	0.72
	4000	0.05					.16	0.97
	8000	0.06					.44	1.00
	15000	0.08					.82	1.00
	20000	0.10					.93	1.00
	30000	0.15	0.78	1.00	30000	0.44	0.99	1.00
Incidence = 0.6 per 1000 Q35.9: cleft palate, unspecified	2000	0.05	0.05	0.05	2000	0.06	0.06	0.16
	4000	0.05	0.05	0.14	4000	0.05	0.07	0.50
	8000	0.05	0.08	0.34	8000	0.06	0.12	0.85
	15000	0.06	0.09	0.77	15000	0.07	0.23	0.99
	20000	0.06	0.11	0.93	20000	0.08	0.31	1.00
	30000	0.06	0.17	1.00	30000	0.10	0.51	1.00

Poisson has greater power than Bernoulli

A minimum of 4000 exposed pregnancies is necessary to observe a doubling in risk of common outcomes with approximately 90% power

Question 2: Can we increase power by using a different propensity score matching method?

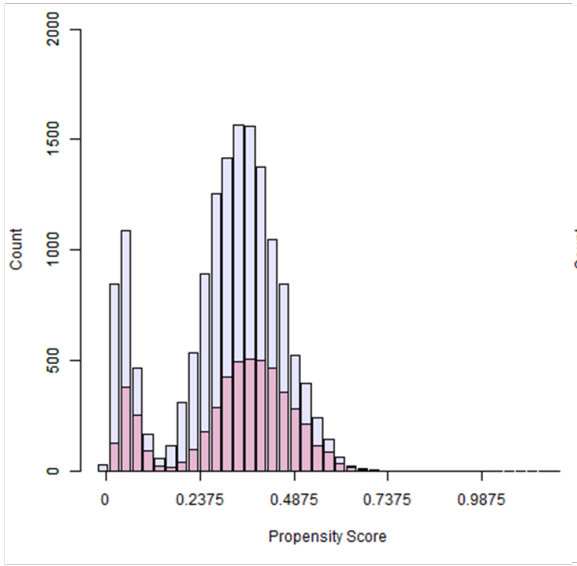
- As we just saw, the Bernoulli model has less power than the Poisson model when we use 1:1 matching
- If we increase the referent to exposed matching ratio to 2:1 or 3:1, will that help increase power?
  - When the referent group is large, including more referent patients may increase power
  - However, fixed ratio matching will exclude exposed patients if there aren't enough referent patients that are close enough for a match, which could decrease power
- We simulated propensity score distributions with varying levels of overlap and calculated power after increasing the fixed matching ratio

Simulated propensity score distributions

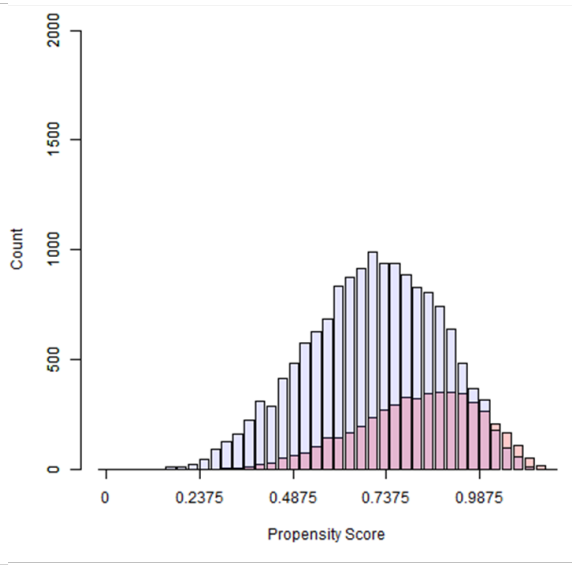
Base population:

- 5,000 exposed pregnancies and 20,000 comparator exposed pregnancies for scenarios A-C
- 5,000 exposed and 495,000 unexposed pregnancies for scenario D

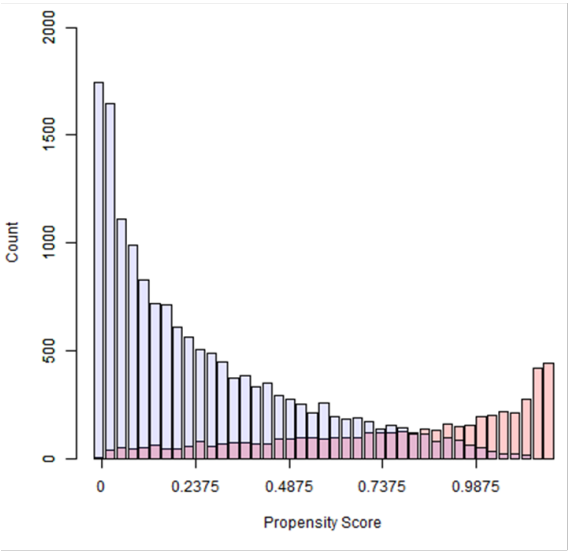
A: active comparator



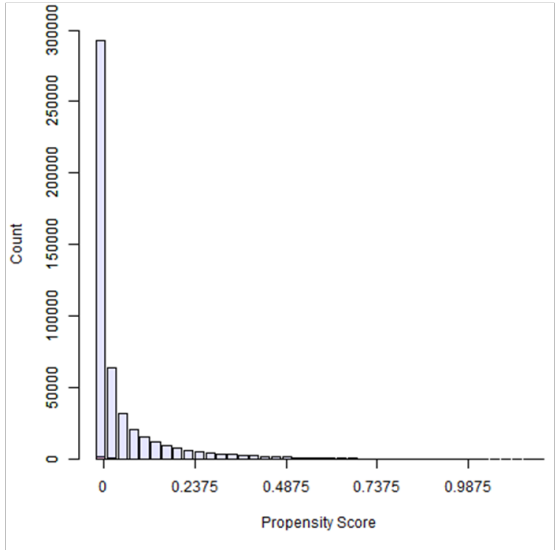
B: active comparator



C: active comparator



D: unexposed comparator



Fixed matching results

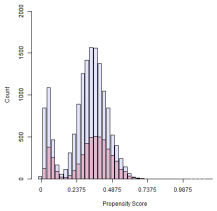
Incidence = 8 per 1000
Q21.0: ventricular septal defect

A

B

C

Active comparator with decreasing overlap in propensity score distributions



D: Unexposed comparator

Matching ratio	Exposed N	Change from 1:1	Referent N	Change from 1:1	Full N	Power for Q21.0 for RR=2
1:1	4,999		4,999		9,998	0.68
2:1	4,997	0%	9,994	100%	14,991	0.81
3:1	4,711	-6%	14,133	183%	18,844	0.84

Fixed matching results

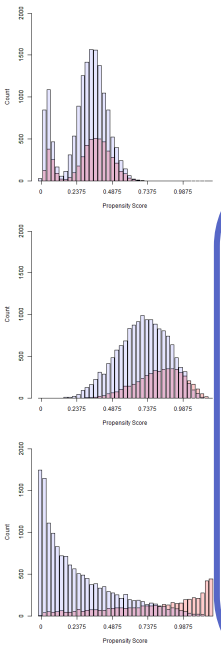
Incidence = 8 per 1000
Q21.0: ventricular septal defect

A

B

C

Active comparator with decreasing overlap in propensity score distributions



D: Unexposed comparator

Matching ratio	Exposed N	Change from 1:1	Referent N	Change from 1:1	Full N	Power for Q21.0 for RR=2
1:1	4,999		4,999		9,998	0.68
2:1	4,997	0%	9,994	100%	14,991	0.81
					344	0.84
					2	0.69
					96	0.75
					44	0.62
					0	0.36
					3	0.34
					16	0.28
1:1	4,892		4,892		9,784	0.68
2:1	4,833	-1.2%	9,666	98%	14,499	0.80
3:1	4,766	-2.6%	14,298	192%	19,064	0.82

Power improved only when:

a) when there was high overlap of propensity scores among groups and enough referent patients

b) when there was an unexposed comparator

Not a reliable method for increasing power

Question 3: How does our outcome definition impact power, given outcome misclassification?

- We used a broad outcome definition to capture all possible events: presence of a single diagnosis code in any care settings that meets the incidence criteria
- This definition is expected to be very sensitive, but probably not very specific

- Low specificity → bias in relative effect estimates, which can reduce our power to detect to true increase in risk
- Low sensitivity → small numbers of observed cases, which also limits power

- Simultaneous scanning of hundreds of events does not allow for targeted outcome definitions that maximize specificity or sensitivity
  - Need a one-size-fits-all outcome definition
- For signal identification, is sensitivity or specificity more impactful on our ability to identify potential alerts?

We can do a bias analysis to address this

- We started with our simulated cohorts from Question 1 and assumed these counts were the true outcome counts, with no misclassification
- We then assumed various combinations of sensitivity and positive predictive value for the selected outcome, and calculated biased outcome counts
  - Specificity is rarely known for outcomes in administrative data, therefore we used positive predictive value instead
- Power was calculated using the biased outcome counts, repeating the methods from Question 1 analyses

Results: Bernoulli model

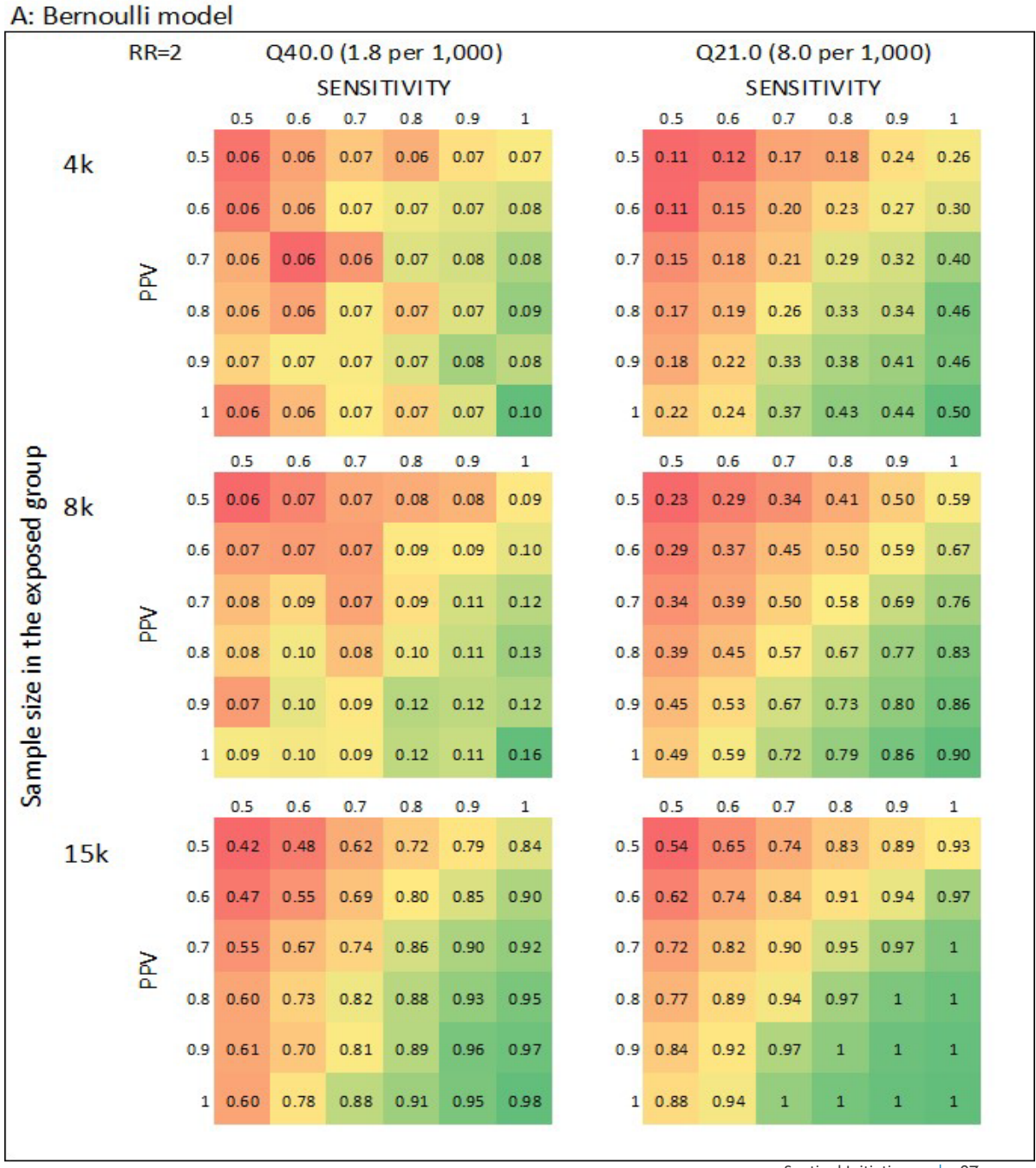
In each square:

Increasing sensitivity

Increasing PPV

Darker green = greater power

Concentrated on the lower right side, where sensitivity is greater than PPV



Results: Poisson model

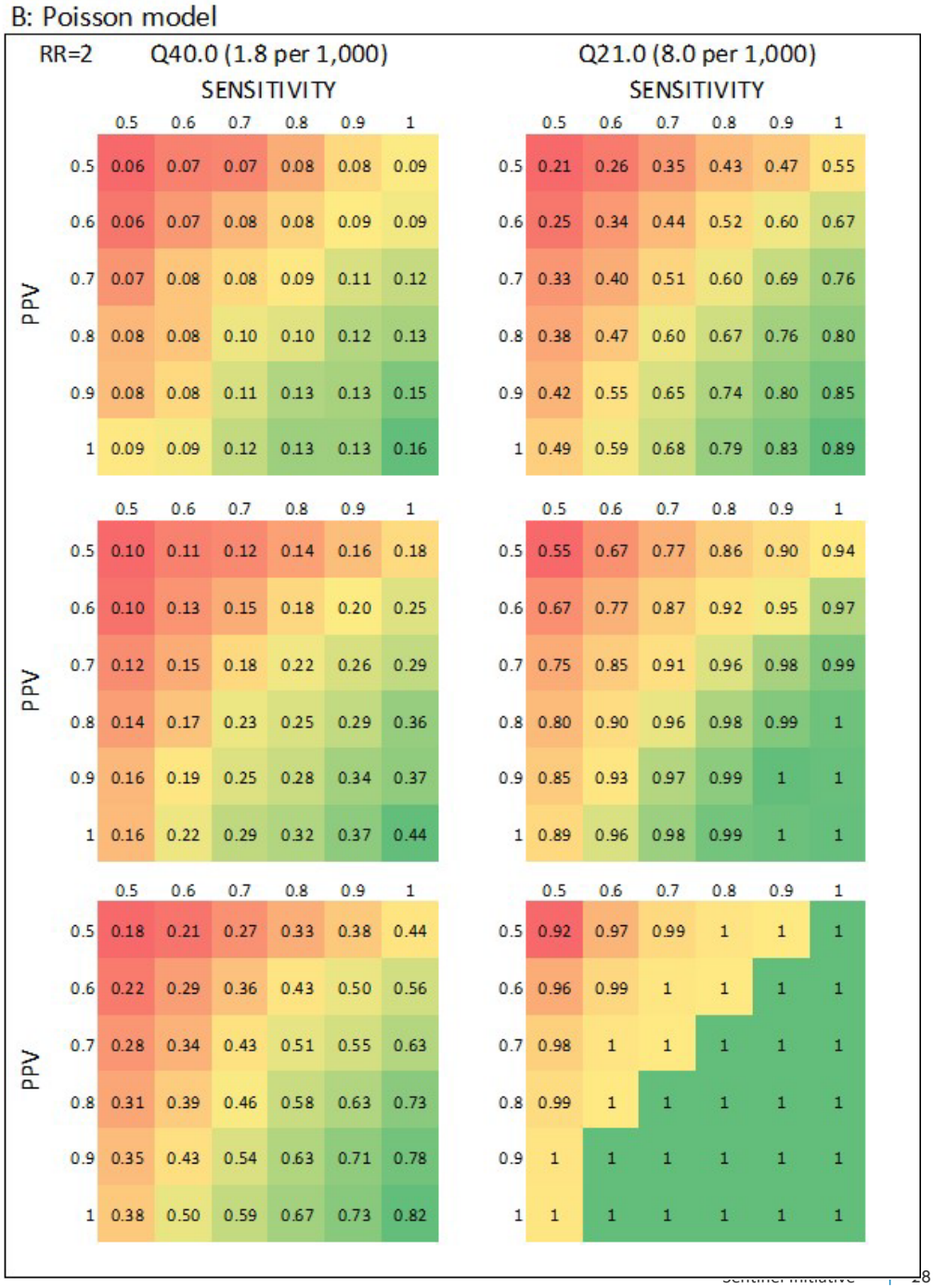
In each square:

Increasing sensitivity

Increasing PPV

Darker green = greater power

Concentrated on the lower right side, where sensitivity is greater than PPV



Results: Poisson model

In each square:

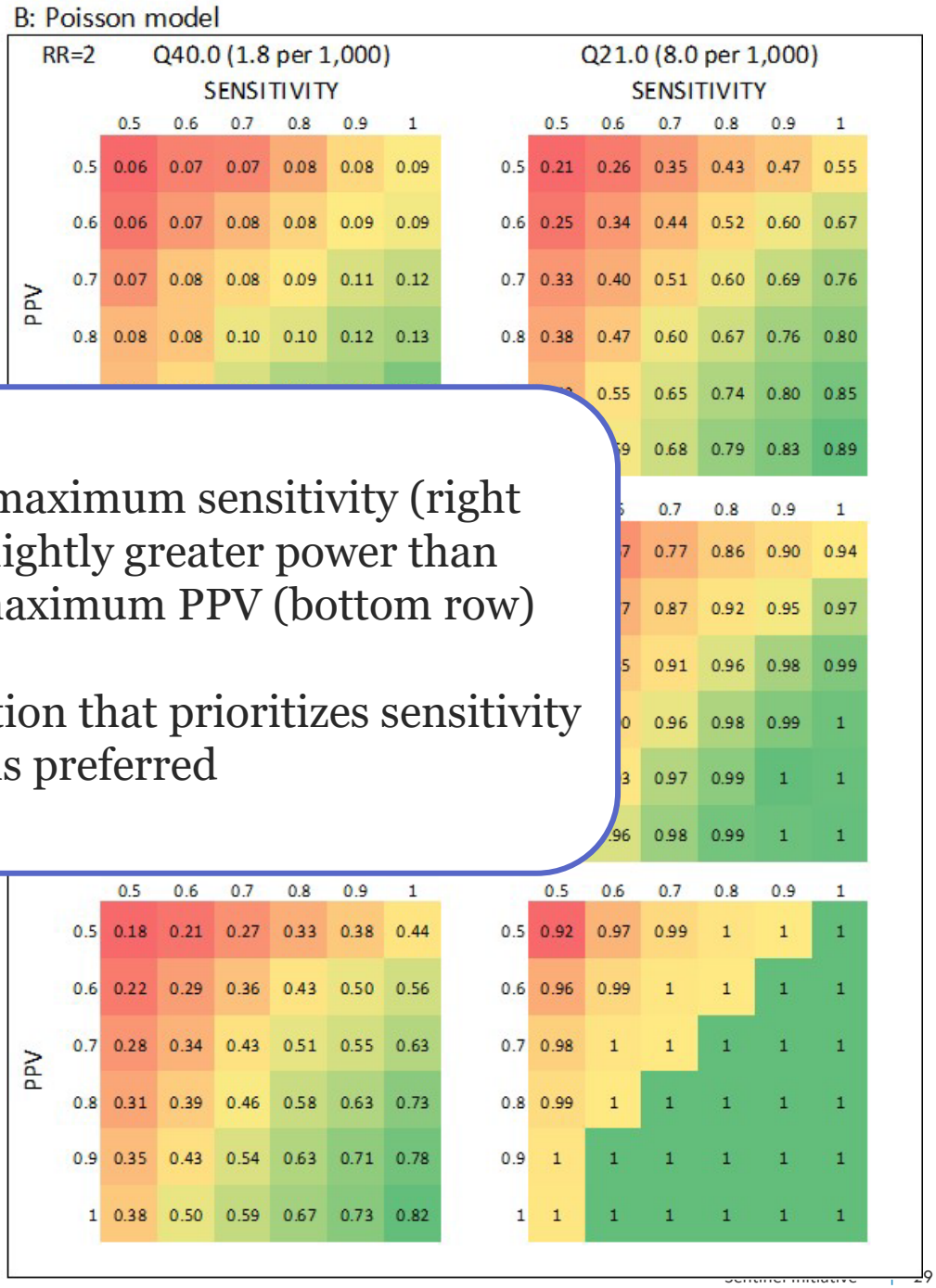
Increasing PPV

Darker green = greater power

Concentrated on the lower right side, where sensitivity is greater than PPV

Scenarios with maximum sensitivity (right column) had slightly greater power than scenarios with maximum PPV (bottom row)

An outcome definition that prioritizes sensitivity is preferred



Simulation conclusions

- We recommend using the Poisson model to increase power to observe alerts
- A potential disadvantage of using the Poisson model is that matching is expected to result in better confounding control than stratification
  - We attempted to improve power using the Bernoulli method by using N:1 fixed ratio matching, but this proved unreliable as a general strategy
- For our purposes, power is more important than confounding control
  - An observed alert can be investigated in a targeted study, where uncontrolled confounding can be mitigated
- Our outcome misclassification bias analysis suggests a highly sensitive outcome definition is useful for maintaining power, regardless of TreeScan model used

Case study

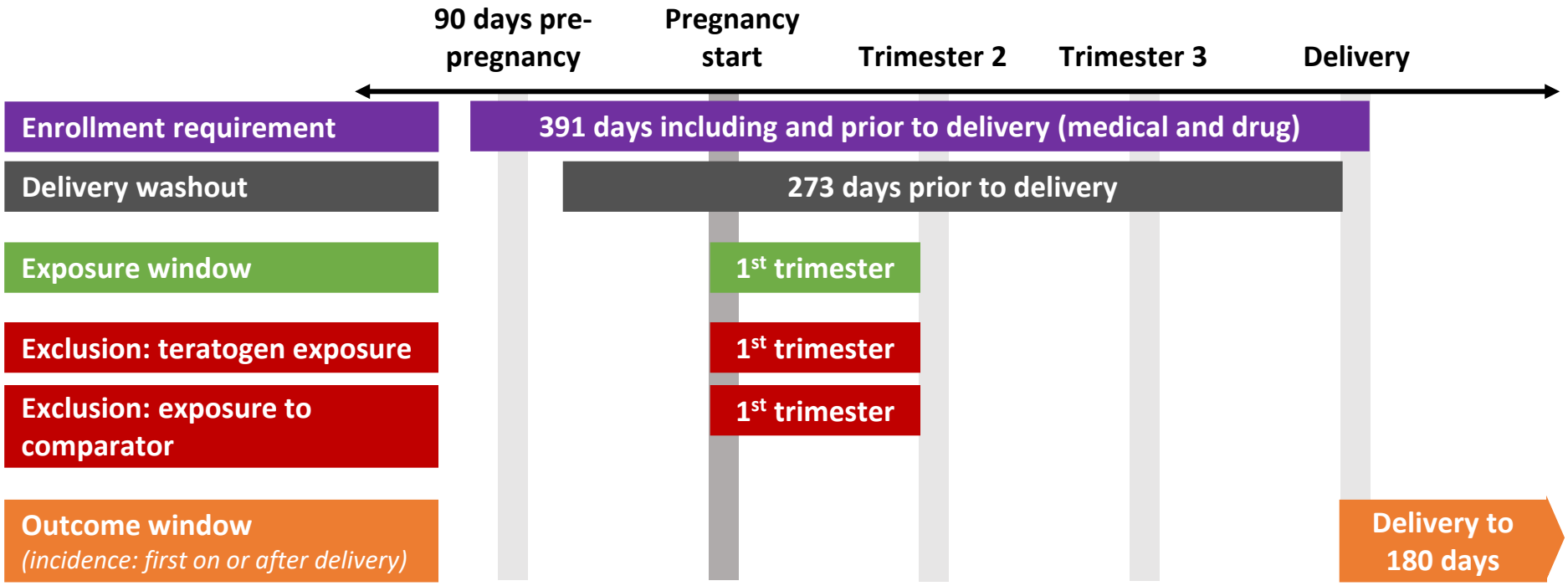
Fluoroquinolones vs Cephalosporins in first trimester

Purpose of the case study

- **Demonstrate the use of TreeScan in real-world data, in a cohort of pregnant women linked to their live-born infants**
- Not designed to identify a new safety risk, therefore we chose drugs with known risk profiles and no known safety issues
  - Expected results: no new alerts
- Selected case study: fluoroquinolone exposure in first trimester compared to cephalosporin exposure in first trimester
  - Antibiotics used to treat a variety of infections in pregnancy

Study design

Data source	IBM MarketScan® Research Database
Eligible population	Women with live birth deliveries between October 1, 2015, and December 31, 2018, aged 10-55 years at delivery



Propensity score models

1. **General model:** selected a general list of variables potentially related to increases in risk of adverse pregnancy outcomes that could be reused in future TreeScan evaluations
   - Similar to previous work to create a general propensity score model for the adult population (Wang, 2021)
   - Included: demographics, pre-existing conditions, screening behaviors, health care utilization metrics
2. **General model + indications:** added indications for fluoroquinolones and cephalosporins
   - Urinary tract and kidney infections, lower respiratory tract infections, ear, nose, and throat infections, gastrointestinal infections, and sexually transmitted infections
3. **High-dimensional propensity score:** used a data driven approach to select variables that are associated with the exposure

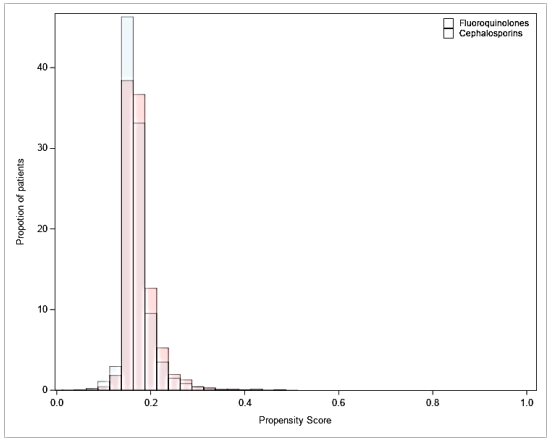
Wang SV, Maro JC, Gagne JJ, Patorno E, Kattinakere S, Stojanovic D, et al. A General Propensity Score for Signal Identification Using Tree-Based Scan Statistics. Am J Epidemiol [Internet]. 2021 Feb 22 [cited 2021 Jun 4]; Available from: https://doi.org/10.1093/aje/kwab034

Select analyses

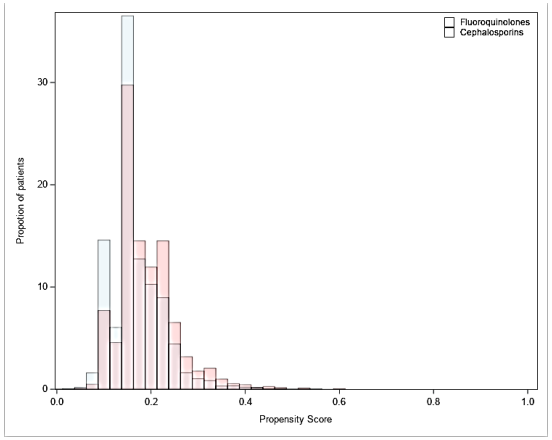
- Propensity score matched design
  - Using the TreeScan Bernoulli model
  - Main analysis: 1:1 matched
  - Sensitivity analyses: 2:1 matched, 3:1 matched
- Propensity score stratified design
  - Using the TreeScan Poisson model
  - Calculated expected counts within deciles of the propensity score
- Other sensitivity analyses varied incidence criteria and outcome definitions and will not be presented – results were consistent with main results

Propensity score distributions

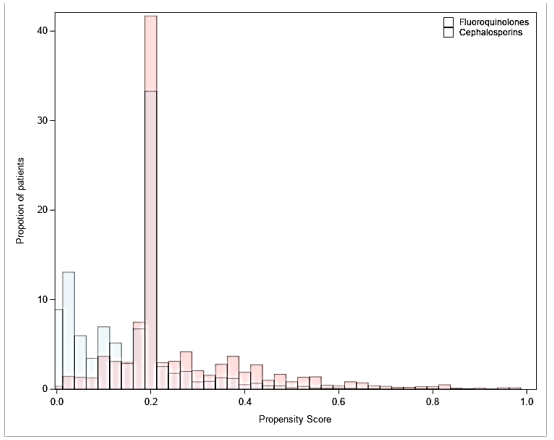
General model



General model + indications



HDPS



- Red = fluoroquinolones, Blue= cephalosporins
- Very good overlap in distributions between the groups in all models
- Adding indications and using HDPS differentiated groups more – potentially better confounding control

Results using propensity score matching and the Bernoulli model

Analysis	Fluoroquinolone exposed N	Fluoroquinolone exposed N cases	Cephalosporin exposed N	Cephalosporin exposed N cases	TreeScan Results
TOTAL	1,791		8,739		
1:1 matched, general model	1,791	504	1,791	494	Q31grp (Congenital malformations of larynx) was significant ($p<0.05$)
1:1 matched, general + indications model	1,790	506	1,790	502	No significant alerts
1:1 matched, HDPS model	1,732	494	1,732	486	No significant alerts
2:1 matched, general + indications model	1,787	510	3,574	1,028	No significant alerts
3:1 matched, general + indications model	1,684	484	5,052	1,448	No significant alerts

Triaging the observed alert: is it worth investigating?

Observed cases:

Code	Description	Fluoroquinolones	Cephalosporins
Q31	Total cases: Congenital malformations of larynx	27	7
Q31.5	Congenital laryngomalacia	25	7
Q31.8	Other congenital malformations of larynx	2	0

- Abnormality of the larynx that leads to collapse of the airway during inspiration
- Clinical presentation:
  - Presents at birth or shortly after, and mild cases resolve by 12-18 months
  - Clinical diagnosis, confirmed with laryngoscopy and bronchoscopy
- Managed expectantly or with acid suppression, speech/swallow therapy and high calorie formula, depending on severity

Triaging the observed alert: is it worth investigating?

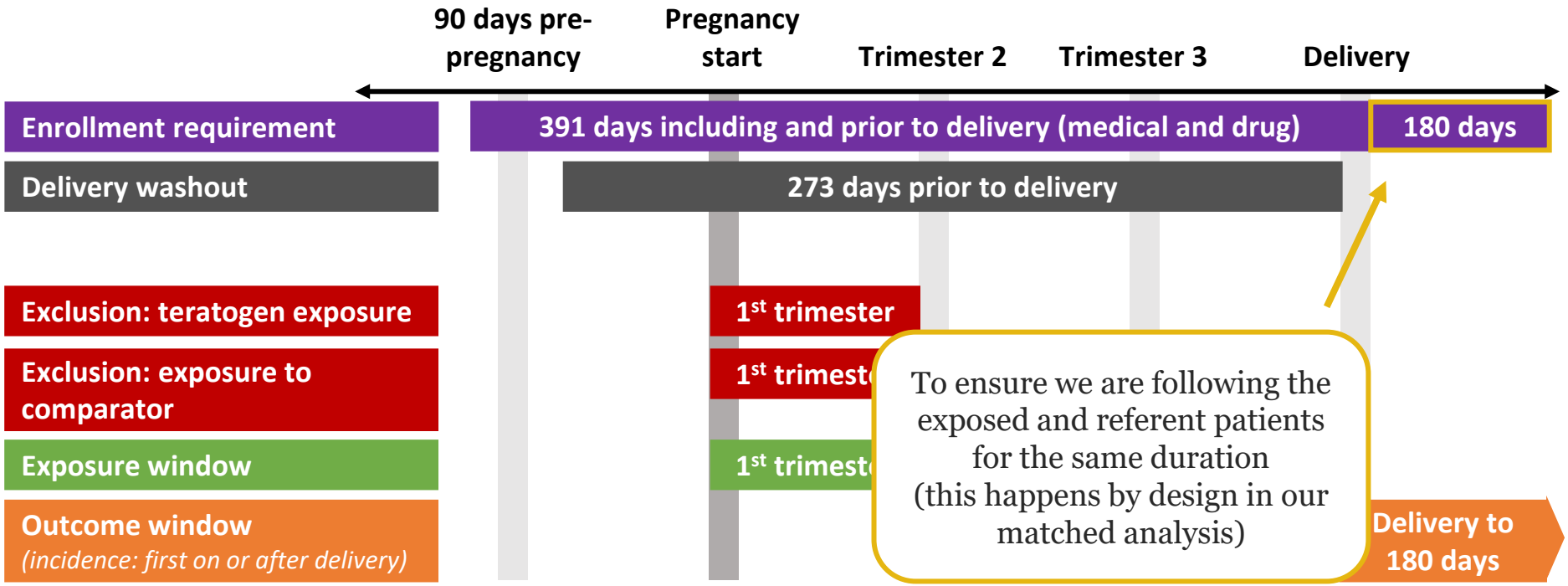
- We provided claims profiles – a list of all maternal and infant claims around of the time of pregnancy and delivery – for all cases for review by FDA workgroup members
- Congenital malformations of the larynx are generally not considered serious and often do not require intervention
- The observed alert was likely due to uncontrolled confounding, given that we did not observe it in analyses with theoretically better confounding control
- Conclusion: no need for additional follow-up

Select analyses

- Propensity score matched design
  - Using the TreeScan Bernoulli model
  - Main analysis: 1:1 matched
  - Sensitivity analyses: 2:1 matched, 3:1 matched
- Propensity score stratified design
  - Using the TreeScan Poisson model
  - Calculated expected counts within deciles of the propensity score
- Other sensitivity analyses varied incidence criteria and outcome definitions and will not be presented – results were consistent with main results

Study design: small change for the stratified analysis

Data source	IBM MarketScan® Research Database
Eligible population	Women with live birth deliveries between October 1, 2015, and December 31, 2018, aged 10-55 years at delivery



Results using propensity score stratification and the Poisson model

Analysis	Fluoroquinolones N	Fluoroquinolones N cases	Cephalosporins N	Cephalosporins N cases	TreeScan Results
Full cohort	1,509		7,165		
Stratified Poisson, general model	1,508	426	7,160	2,030	Q513grp and Q513ngrp: bicornate uterus
Stratified Poisson, general + indications	1,507	426	7,155	2,028	Q513grp and Q513ngrp: bicornate uterus
Stratified Poisson, HDPS	1,500	423	7,089	2,008	Q513grp and Q513ngrp: bicornate uterus

- This is very likely associated with the mother's record
  - We include outcomes recorded in the mother's or infant's record after delivery because the infant may have a 30-60 day gap between delivery and insurance enrollment
  - This may result in false alerts like we observe here, but they are easily explained and individual maternal and infant records can be reviewed to confirm

Why did we see different results by method?

- The Poisson model has greater power than the Bernoulli model, therefore alerts observed with Poisson may not be able to be observed using Bernoulli
- Different propensity score methods result in slight changes to the referent population, resulting in different expected counts
  - The alert observed in the 1:1 matched analysis using the general propensity score model likely resulted in very tight control using a mis-specified model
  - Adding indications or using HDPS resulted in no alerts in the matched analysis

Summary of the empirical study results

- We did not observe evidence that fluoroquinolone use in first trimester increases the risk of adverse infant outcomes when compared to cephalosporin use in first trimester
- At 1791 fluoroquinolone exposed, we are underpowered to see smaller increases in risk (this is supported by the simulation results)
- Use of propensity score stratification did not result in many spurious alerts
  - In this active comparator setting, a slight decrease in confounding control is likely worth the increase in power attained by using Poisson vs Bernoulli

Conclusions

Conclusions

- TreeScan is a promising method for use in surveillance of potential adverse infant events following maternal medication exposure during pregnancy
- If less than 4000 exposed pregnancies are available for study, the analysis may be underpowered to detect most alerts
- Using TreeScan in administrative data within Sentinel offers notable advantages:
  - Utilize the large sample sizes available in administrative data, and build off previous methods to identify pregnancies and pregnancy exposures
  - Not limited to major congenital malformations as a primary outcome – can scan for all types of malformations individually and in clinically relevant groupings (e.g., atrial septal defect, any cardiac malformation)
- Results on appropriate methods and utility of using TreeScan for adverse maternal outcomes are forthcoming

Questions?

Contact: Elizabeth_suarez@harvardpilgrim.org