Evaluating Use of Methods for Adverse Event Under Surveillance (EUMAEUS)
Why EUMAEUS?

1) The rapid rollout of COVID-19 vaccines makes it increasingly critical to perform large-scale evaluations of vaccine safety using real-world evidence.

2) Estimate the comparative performance (bias, precision, timeliness) of the case-control, cohort, historical rate, and self-controlled methods for vaccine safety.
Literature Review

Lana Lai

on behalf of the EUMAEUS task force
# Types of Study Designs

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Clinical Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>• Comparison of cases vs. non-cases from the same source population from the same time-period</td>
<td>• Uses small data sample from entire cohort, cost efficient • Use matching to control for time-varying confounders</td>
<td>• Confounding by indication • Selection bias • Misclassification of exposure</td>
<td>• Autism spectrum disorders &amp; various vaccines • Inflammatory bowel disease (IBD) &amp; MMR vaccine • Guillain-Barré syndrome (GBS) &amp; H1N1 vaccine</td>
</tr>
</tbody>
</table>
## Types of Study Designs

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Clinical Applications</th>
</tr>
</thead>
</table>
| **Cohort**           | • Comparison of incidence rate ratio of adverse events between vaccinated vs. unvaccinated population | • Easy to implement – large amount of data available  
  • Use matching / stratification to control for potential confounders | • Confounding by indication  
  • Misclassification of exposure | • Intussusception & rotavirus vaccine  
  • Autism spectrum disorders & various vaccines |
| **Historical Rate (Comparator) Cohort** | • Comparison between observed incidence of adverse events vs. expected incidence based on historical data | • Greater statistical power to detect rare adverse events  
  • Improved timeliness in detecting potential safety signals | • Temporal confounders (e.g. seasonality, changing trends in detection of adverse events & variation in diagnostic criteria over time) | • Pediatric vaccines  
  • Tdap vaccine  
  • HPV vaccine  
  • H1N1 vaccine |
# Types of Study Designs

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Clinical Applications</th>
</tr>
</thead>
</table>
| Self-Controlled Case Series  | Comparison between incidence rates in exposed time periods vs. incidence rates of self-matched unexposed time periods | • Adjust for time-invariant confounders  
• SCCS: Multiple occurrences of independent events within an individual can be assessed  
• SCRI: Less susceptible to misclassification of exposure | • Time-varying confounding (e.g. age, seasonality)  
• Reverse causality bias                                                                 | • Guillain-Barré syndrome (GBS) & H1N1 vaccine  
• Autism spectrum disorders & various vaccines  
• Seizures & various vaccines                                                                 |
Overview of the EUMAEUS Experiment Design

Marc Suchard

on behalf of the EUMAEUS task force
EUMAUES is an empirical benchmark study

Builds on our prior work evaluation of comparative (drug) effectiveness and safety methods published in *Harvard Data Science Review*

To systematically evaluate the performance of methods to reliably identify vaccine safety signals in real-world settings.
Vaccine safety surveillance methods

Reduce systematically to four components:

• Construction of a \textit{counterfactual} (“expected count” without vaccination)

• A \textit{time-at-risk} when safety events can occur

• The \textit{test-statistic} to estimate, and

• A \textit{decision rule} to classify signals from non-signals
Counterfactual construction

• Case-control
  – How often are patients with events vaccinated?

• Contemporary non-user comparator cohort method
  – How often do events occur to similar unvaccinated patients?
  – Some variants: anchoring (or not) on healthcare visit; matching (or not) on age + sex

• Historical rates
  – How often did events occur to other patients in the past?
  – Some variants: anchoring; stratifying (or not) on age + sex

• Self-control case series
  – How often did/do events occur in the same patients at different times?

Note: 17 total variations drawn from the literature
Time at risk and test-statistics

• A **time-at-risk** when safety events can occur:
  – 0-1 days, **1-28 days** and 1-42 days after vaccination
  – Dose definition (first, second, both)

• The **test-statistic** to estimate:
  – Effect-size estimation (incidence rate ratio, hazard ratio or odds ratio)
  – Log-likelihood ratio (common in vaccine surveillance, allows for corrections for multiple testing over time via MaxSPRT)
  – With and without **empirical calibration** (to control for systematic error)
Method performance metrics

- A **decision rule** to classify signals from non-signals
  - Bias / variance (particularly of the *residual systematic error*)
  - Type 1 error rate
  - Type 2 error rate
  - *Timeliness* to achieve power

---

**IDEAL:** 5% (nominally) → 0% (100% power)
Real-world evidence with 117M estimates

Exposures of interest:
- H1N1pdm (‘09-‘10)
- Seasonal influenza (Fluvirin, ‘17-‘18)
- Seasonal influenza (Fluzone, ‘17-‘18)
- Seasonal influenza (all, ‘17-‘18)
- Zoster (2018, 2 doses)
- HPV (2018, 2 doses)

Data sources:
- CCAE
- MDCR
- MDCD
- Optum EHR

Negative control outcomes (93):
- Not related to any of these vaccines
- Similar prevalence and %-inpatient diagnoses (severity) to AESI
- Clinical expert review

Positive control outcomes:
- Imputed from negative controls
- Known effect sizes (1.5, 2, 4 x)

Open Science: pre-specified and registered protocol, open-source analytic code, public access to all results

- [https://github.com/ohdsi-studies/Eumaeus/](https://github.com/ohdsi-studies/Eumaeus/)
- [https://data.ohdsi.org/Eumaeus/](https://data.ohdsi.org/Eumaeus/)
Prelude to the results

• Which methods are least bias in the real-world?
  – Effect of counterfactual anchoring
  – Effect of confounding adjustment

• What is the trade-off to achieve, say, 50% power?

• Should we combine multiple designs (signal generation / evaluation) to improve performance?

• Is sequential testing (α-spending) correction a panacea?

• Do 2nd doses influence method choice?
Bias, precision and timeliness of historical rate comparison methods

Xintong Li

on behalf of the EUMAEUS task force
Recall the advantages of historical comparator design:
   – Greater statistical power
   – Improved timeliness

Especially useful at early stage after vaccine introduction

Historical comparator is from:
   – literature
   – within same database / population (best-case scenario)
   – others
# Choice of design

<table>
<thead>
<tr>
<th>Population</th>
<th>Time-at-risk</th>
<th>Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>Entire year</td>
<td>Yes</td>
</tr>
<tr>
<td>Adjusted for age and sex</td>
<td>Relative to outpatient visit</td>
<td>No</td>
</tr>
</tbody>
</table>
Historical comparison in general: Sensitive but not specific
Sensitive but not specific

Adjust for age and sex reduced type 1 error.

Unadjusted, entire historical period

Age and sex adjusted, entire historical period
Sensitive but not specific

After adjusting for age and sex, anchoring on visit further reduce type 1 error.
Empirical calibration "forcing" type 1 to close to nominal, at the cost of increasing type 2 error.
Empirical calibration: reduce type 1, increase type 2
Higher and faster uptake, earlier detection

Database
CCAE

Analysis
Adjusted for age and sex, no anchoring

Calibration
No
Conclusion

• Sensitive but not specific: overestimate risks
• Age-sex adjustment reduce false positive
• Anchoring on visit reduce false positive
• Empirical calibration: forced type 1 error back to normal, at the cost of increasing type 2 error.
• For vaccine with high uptake speed: can detect earlier, stabilized estimation.
Combining Methods in a Safety Surveillance System

Faaizah Arshad

on behalf of the EUMAEUS task force
HIV testing

- Two part test: 1) highly sensitive (few false negatives); 2) highly specific (eliminate false positives)
Methods

• We hypothesized that sequentially combining methods might be desirable for population-level COVID-19 vaccine safety surveillance.

• Method 1: historical comparator (sensitive / cheap)
• Method 2: self-controlled case series (specific)
• Combined: Method 1 → Method 2
Uncalibrated type I and II errors for all outcomes

Database: IBM MDCR
Outcome: All
Time-at-Risk: 1-28 days

SCCS shows lower type I error (fewer false positives); higher specificity.

Historical comparator remains more sensitive than SCCS. Lower type II error and fewer false negatives.

Combined method shows low type I error and greater type II error (specific but not sensitive).
Calibrated type I and II errors for all outcomes

**Database:** IBM MDCR  
**Outcome:** All  
**Time-at-Risk:** 1-28 days

Calibration tries to fix the type I error rate (closer to nominal); most noticeable for historical comparator.

After calibration, historical comparator still most sensitive. Reduced type I error for historical comparator and SCCS.
Conclusion

• Reject hypothesis.

• Sequentially combining sensitive and specific methods does not improve performance over using a single method.

• Future vaccine monitoring should consider the sequence of methods used to ensure accurate signal detection.
Estimation for Two-Dose Vaccines

Ty Stanford
on behalf of the EUMAEUS task force
Aim:
• Does the inclusion of data from the 2nd dose, among vaccines with 2 doses, reduce type II error?

Data:
• This limits EUMAEUS data to
  (CCAE, Optum EHR) x (HPV vaccine, Zoster vaccine) combinations
Dose accumulation

<table>
<thead>
<tr>
<th>Database</th>
<th>Dose</th>
<th>HPV vaccination (Gardasil 9)</th>
<th>Zoster vaccination (Shingrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optum EHR</td>
<td>First</td>
<td>233985</td>
<td>219665</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>28336</td>
<td>63464</td>
</tr>
</tbody>
</table>
To calibrate or not to calibrate?

Uncalibrated

Calibrated

Adj. for sequential testing

MaxSPRT

Dose 1st only
Adding 2nd dose: Cohort Design

Uncalibrated

Calibrated

Dose
(1st only) vs (1st & 2nd)
Adding 2nd dose: SCCS

Uncalibrated vs Calibrated

Dose: (1st only) vs (1st & 2nd)

Adj. for sequential testing
Conclusion

• Inclusion of the 2\textsuperscript{nd} dose can increase the power
  – marginally in this case, likely as a result of a marginal increase in sample size

• The most important factor is \textit{empirical calibration}
  – more data doesn’t magically negate issues with specific designs

• Future work to understand the issues better:
  – Larger proportion of 1\textsuperscript{st} doses to also have 2\textsuperscript{nd} doses (with differing rates)
  – Underlying signals (positive controls) to have varying effects after each dose
Comparison of performance across methods

Martijn Schuemie

on behalf of the EUMAEUS task force
Same data & question, different methods: different results

<table>
<thead>
<tr>
<th>Analysis choices</th>
<th>Effect size (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched case-control</td>
<td>3.33 (0.64-15.44)</td>
</tr>
<tr>
<td>PS-weighted cohort method</td>
<td>4.42 (1.83-10.42)</td>
</tr>
<tr>
<td>Historical comparator</td>
<td>3.50 (1.92-5.87)</td>
</tr>
<tr>
<td>SCCS</td>
<td>1.07 (0.59-1.81)</td>
</tr>
</tbody>
</table>

Exposure
H1N1pdm vaccinations

Outcome
Contusion of toe

Database
Optum EHR
Comparing on type 1 and type 2 error

Case-control and historical comparator tend to generate many false positives
Empirical calibration: restoring type 1

Calibration makes methods comparable.

After calibration (fixed type 1), SCCS has lowest type 2 error.

Database
Optum EHR
## Adjusting for systematic error and sequential testing

<table>
<thead>
<tr>
<th>Type 1 error</th>
<th>Historical comparator</th>
<th>SCCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncalibrated, no adjustment for sequential testing</td>
<td>28.0%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Uncalibrated, MaxSPRT</td>
<td>18.3%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Calibrated, no adjustment for sequential testing</td>
<td>10.8%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Calibrated, MaxSPRT</td>
<td>6.5%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

**Exposure**
- H1N1pdm vaccinations

**Outcome**
- All negative controls

**Database**
- Optum EHR

Adjusting for **systematic error** has bigger impact than adjusting for **sequential testing**.
Time to 50% sensitivity (after calibration)

- Matched case-control
- PS-weighted cohort method
- Historical comparator
- SCCS

SCCS has shortest time to detection

All methods struggle to achieve 50% sensitivity for small effects

Adj. for sequential testing
MaxSPRT

Adj. for systematic error
Empirical calibration

Database
Optum EHR
More or less consistent across methods / outcomes /databases

<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
<th>H1N1 vaccination</th>
<th>HPV vaccination (Gardasil 9)</th>
<th>Seasonal flu vaccination (Fluquin)</th>
<th>Seasonal flu vaccination (Fluzone)</th>
<th>Seasonal flu vaccination (Shingrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Adj. for sequential testing</td>
<td>Age &amp; sex matched controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort method</td>
<td>Adj. for systematic error</td>
<td>Age &amp; sex adjusted, using random controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical calibration</td>
<td>Unadjusted, using outpatient visits as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS matching, using outpatient visits as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unadjusted, using random days as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS matching, using random days as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS stratification, using outpatient visits as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS stratification, using random days as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS weighting, using outpatient visits as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS weighting, using random days as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Per-month PS matching, using outpatient visits as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Per-month PS matching, using random days as comparator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical comparator</td>
<td>Adj. for sequential testing</td>
<td>Unadjusted, using entire historic period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adj. for systematic error</td>
<td>Age &amp; sex adjusted, using entire historic period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical calibration</td>
<td>Unadjusted, using TaR after historic visit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age &amp; sex adjusted, using TaR after historic visit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unadjusted, using entire historic period, filtered</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age &amp; sex adjusted, using entire historic period, filtered</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unadjusted, using TaR after historic visit, filtered</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age &amp; sex adjusted, using TaR after historic visit, filtered</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCCS / SCRI</td>
<td>Adj. for sequential testing</td>
<td>Unadjusted SCCS excluding pre-vaccination window</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adj. for systematic error</td>
<td>Age &amp; season adjusted SCCS excluding pre-vaccination window</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical calibration</td>
<td>SCRI with prior control interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCRI with posterior control interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unadjusted SCCS excluding all pre-vaccination time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time to 50% sensitivity (months)

- True effect size = 1.5
- True effect size = 2
- True effect size = 4
Conclusions

• Many methods show large systematic error / type 1 error

• Empirical calibration can restore type 1 error to nominal, at the cost of increasing type 2 error
  (depending on magnitude of systematic error)

• Empirical calibration often has bigger impact than adjusting for sequential testing
  (should do both)

• After calibration and adj. for sequential testing SCCS seems overall best
  (shortest time to detection)

• No method achieves high sensitivity for small true effect sizes
  (on these data)
Recommendations for a safety surveillance system

Martijn Schuemie
on behalf of the EUMAEUS task force
Recommendations

• Many methods (e.g. case-control & historical comparator) have positive bias, causing many false positives (high type 1 error)
  – Include negative controls and use empirical calibration
  – Include self-controlled designs
  – Always use confounding adjustment
  – Carefully consider anchoring of counterfactual

• Detecting more than half of true adverse effects may require accepting more false positives (e.g. using calibrated $p < 0.10$)

• Combining multiple designs likely doesn’t improve performance
  – Do not distinguish between ‘signal generation' and 'signal evaluation'

• Second dose often underpowered to contribute to evidence