The journey toward Population-level Effect Estimation

Martijn Schuemie, PhD
Janssen Research and Development
Population-level effect estimation

• What is the effect of treatment A on outcome X?

• What is the effect of treatment A on outcome X, compared to exposure B?
Population-level effect estimation

- **Evidence Generation**
  - How to produce evidence from the data?

- **Evidence Evaluation**
  - How do we know the evidence is reliable?

- **Evidence Dissemination**
  - How do we share evidence to inform decision making?
Doctor, I'm starting on duloxetine, should I be worried about stroke?

Let me see what I find in the literature...
Evidence from literature

Paper by Lee et al, 2016
- Compare new users of SNRIs (includes duloxetine) vs SSRIs
- Taiwanese insurance claims data
- 12 month washout
- remove people using both drugs
- remove people with a prior history of head injury
- remove people with a prior history of stroke or intracranial hemorrhage
- Propensity score: logistic regression with treatment as dependent variable
- HOI is Stroke: first hospitalization with ICD-9 433, 434, or 436
- time-varying Cox regression using 5 PS strata

<table>
<thead>
<tr>
<th>Main analyses</th>
<th>Crude Hazard Ratio (95% CI)</th>
<th>Adjusted Hazard Ratioa (95% CI)</th>
<th>P</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRIs (n = 76,920) vs SSRIs (n = 582,650)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>0.92 (0.83–1.02)</td>
<td>1.01 (0.90–1.12)</td>
<td>.12</td>
<td>.91</td>
</tr>
</tbody>
</table>
How reliable is this evidence?

- Can the results be reproduced?
- Did the analysis program do what it was supposed to do?
- Is the estimate unbiased?
- Does the p-value have nominal characteristics?
- Does the confidence interval really represent the uncertainty about the effect size?

Are we really 95% confident the true effect size is between 0.90 and 1.12?
Population-level effect estimation

- How to produce evidence from the data?
‘Replicating’ Lee et al.

Our replication:

• Compare new users of **Duloxetine (SNRI) vs. Sertraline (SSRI)**
• **US insurance claims data (Truven CCAE)**
• 12 month washout
• remove people using both drugs
• remove people with a prior history of stroke
• **restricted to people with a diagnosis of major depressive disorder and no prior diagnosis of bipolar disorder or schizophrenia**
• Propensity score: **regularized** logistic regression with treatment as dependent variable, and **used 58,285 covariates**
• **HOI is Stroke:** first hospitalization with ICD-9 433,434, or 436 (but then coded as standard concepts)
• **fixed-time** Cox regression using **10 PS strata**
OHDSI recommendations for evidence generation

- Post protocol online
  - Prespecify research objectives and design decisions

- Make study code open source
  - From CDM to hazard ratios

- Use validated software
  - OHDSI Methods Library uses unit tests and simulation

- Replicate across several databases
  - 4 included so far, more will follow

https://github.com/OHDSI/StudyProtocols/LargeScalePopEst
Population-level effect estimation

- How do we know the evidence is reliable?
Standard diagnostics

Most study designs have diagnostics that could be used, e.g.

• Propensity score distribution overlap

• Covariate balance
Diagnose the propensity score distribution

Duloxetine vs. Sertraline

Only 45% of patients are near clinical equipoise, most patients are systematically different from the comparator group.

We therefore know crude analysis will likely be biased.

Any covariate adjustment strategy that corrects for this bias will result in impact in the generalizability of the findings to the original research question.

Results from Truven CCAE

Duloxetine: n = 90,043
Sertraline: n = 175,950
Diagnose covariate balance

Standardized difference of mean

Age group 10-14

<table>
<thead>
<tr>
<th></th>
<th>Duloxetine</th>
<th>Sertraline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before stratification</td>
<td>0.2%</td>
<td>3.8%</td>
</tr>
<tr>
<td>After stratification</td>
<td>0.3%</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

After stratification on the propensity score, all 58,285 covariates have standardized difference of mean < 0.1
Empirical evaluation of the study

• Control
  exposure-outcome for which the effect size is known

• Negative control
  exposure-outcome where relative risk is believed to be 1

• Negative controls for comparative effectiveness
  outcomes not believed to be caused by either treatments

Example: ingrowing nail
Negative control: ingrowing nail

Crude estimate:

HR = 1.16 (1.01 – 1.32), p = 0.03
Negative control: ingrowing nail

Not statistically significant
Negative control: ingrowing nail

Statistically significant
Negative control: ingrowing nail

Adjusted estimate:

HR = 0.94 (0.80 – 1.10), p = 0.44
<table>
<thead>
<tr>
<th>Condition</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acariasis</td>
<td>Ingrowing nail</td>
</tr>
<tr>
<td>Amyloidosis</td>
<td>Iridocyclitis</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>Irritable bowel syndrome</td>
</tr>
<tr>
<td>Aseptic necrosis of bone</td>
<td>Lesion of cervix</td>
</tr>
<tr>
<td>Astigmatism</td>
<td>Lyme disease</td>
</tr>
<tr>
<td>Bell’s palsy</td>
<td>Malignant neoplasm of endocrine gland</td>
</tr>
<tr>
<td>Benign epithelial neoplasm of skin</td>
<td>Mononeuropathy</td>
</tr>
<tr>
<td>Chalazion</td>
<td>Onychomycosis</td>
</tr>
<tr>
<td>Chondromalacia</td>
<td>Osteochondropathy</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>Paraplegia</td>
</tr>
<tr>
<td>Croup</td>
<td>Polyp of intestine</td>
</tr>
<tr>
<td>Diabetic oculopathy</td>
<td>Presbyopia</td>
</tr>
<tr>
<td>Endocarditis</td>
<td>Pulmonary tuberculosis</td>
</tr>
<tr>
<td>Endometrial hyperplasia</td>
<td>Rectal mass</td>
</tr>
<tr>
<td>Enthesopathy</td>
<td>Sarcoidosis</td>
</tr>
<tr>
<td>Epicondylitis</td>
<td>Scar</td>
</tr>
<tr>
<td>Epstein-Barr virus disease</td>
<td>Seborrheic keratosis</td>
</tr>
</tbody>
</table>

Generated with the help of LAERTES (see posters)
We would expect 5% of negative controls to have $p < 0.05$

Instead, 68% has $p < 0.05$!

We found crude estimates to be uninformative. Do not use for decision making!
All negative controls - adjusted

When using the propensity score, 16% have $p < 0.05$

In the past, we’ve shown you how you can perform p-value calibration:

- P-value represents probability of estimate when true RR = 1
- Negative controls provide empirical distribution of estimates when RR = 1
- Use empirical null distribution to compute calibrated p-value
After calibration, 4% have $p < 0.05$ (was 16%)

What if HR $<> 1$?

Calibrated $p < 0.05$
Trouble with positive controls

- Often very few positive examples for a particular comparison
- Exact effect size never known with certainty (and depends on population)
- Doctors also know they’re positive, and will change behavior accordingly
Creating positive controls

- Start with negative controls: RR = 1
- Add simulated outcomes during exposure until desired RR is achieved
- Injected outcomes should behave like ‘real’ outcomes: preserve confounding structure by injecting outcomes for people at high risk
Creating positive controls

Patient 1: Duloxetine
Patient 2: Sertraline
Patient 3: Duloxetine
Patient 4: Sertraline
Patient 5: Duloxetine
Patient 6: Sertraline

Predictive model of outcome indicates this is a high-risk patient

New RR = 2 (but with same confounding)
Estimated effects for positive controls

Black line indicates true hazard ratio
Estimating effects for positive controls

Ingrowing nail
True RR = 1
Estimated RR = 0.94 (0.80 – 1.10)
Estimating effects for positive controls

Ingrowing nail+
True RR = 1.5
Estimated RR = 1.47 (1.27 – 1.69)
Estimating effects for positive controls

Ingrowing nail++
True RR = 2
Estimated RR = 1.91 (1.67 – 2.19)
Estimating effects for positive controls

Ingrowing nail+++  
True RR = 4  
Estimated RR = 3.89 (3.53 – 4.48)
Estimating effects for positive controls

Analysis suggests bias remains constant with effect size
Evaluating coverage of the CI

Coverage

83%

Coverage of 83% means the true effect size is outside of the 95% confidence interval 17% of the time (when the true RR = 1)

Coverage decreases with true effect size

70%

Missing the true effect size 30% of the time when the true RR = 2!
Confidence interval calibration

For $HR_{true} = 1$:

$$
\mu = \alpha \mu + \beta \mu \log(HR_{true})
$$

$$
\sigma = \alpha \sigma + \beta \sigma \log(HR_{true})
$$

For $HR_{true} = 2$:

$$
\mu
$$

$$
\sigma
$$
Calibrating a confidence interval

Uncalibrated

0.94 (0.80 - 1.10)

Calibrated

0.90 (0.75 - 1.11)

Confidence intervals were too narrow, so made wider to get to nominal coverage

\[
\begin{align*}
\mu &= 0.04 + 1.01 \log(HR_{\text{true}}) \\
\sigma &= 0.07 + 0.05 \log(HR_{\text{true}})
\end{align*}
\]
Confidence interval calibration

Uncalibrated

Calibrated
Confidence interval calibration

Coverage

96%

91%

91%

Confidence interval calibration complements p-value calibration
#### Current evidence for stroke

Result from Lee et al.

<table>
<thead>
<tr>
<th>Main analyses</th>
<th>Crude Hazard Ratio (95% CI)</th>
<th>P</th>
<th>Adjusted Hazard Ratio$^a$ (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRIs (n = 76,920) vs SSRIs (n = 582,650)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>0.92 (0.83–1.02)</td>
<td>.12</td>
<td>1.01 (0.90–1.12)</td>
<td>.91</td>
</tr>
</tbody>
</table>
Proposed evidence for stroke

Duloxetine vs. Sertraline

Results are comparable to Lee et al., but we provide the context to interpret the results.
OHDSI recommendations for evidence evaluation

✓ Produce standard diagnostics
  • E.g. for cohort studies diagnose the propensity score distribution, covariate balance, etc.

✓ Include negative controls
  • Estimate the error when the null is true

✓ Create positive controls
  • Estimate the error when RR > 1

✓ Calibrate p-value and confidence intervals
  • Restoring nominal characteristics
Population-level effect estimation

- Evidence Generation
- Evidence Evaluation
- Evidence Dissemination

• How do we share evidence to inform decision making?
Evidence dissemination

• Traditionally, this evidence is disseminated through the scientific literature

• How well does that work?
**Automated extraction of effect sizes from literature**

**RESULTS:** In comparison with distant past users of BP, current users of BP showed an almost twofold increased risk of AF: odds ratio (OR) = 1.78 and 95% CI = 1.46-2.16. Specifically, alendronate users were mostly associated with AF as compared with distant past use of BP (OR, 1.97; 95% CI 1.50-2.43).
Observational research results in literature

85% of exposure-outcome pairs have $p < 0.05$

29,982 estimates
11,758 papers
What went wrong?

- Observational study bias
- Publication bias
- P-hacking
I have a headache and my stomach really hurts!

I'll prescribe drug A for your headache, it's safe for people at risk of stomach bleeding.

One week later...

I took drug A, now I have a stomach bleeding!

Ha! Drug A causes stomach bleedings!
Publication bias

---

**Jelly beans cause acne!**

Scientists! Investigate!

But we're playing Minecraft!

...Fine.

---

We found no link between jelly beans and acne ($P > 0.05$).

That settles that!

I hear it's only a certain color that causes it.

Scientists!

But Minecraft!

---

News report:

Green jelly beans linked to acne!

95% confidence

Only 5% chance of coincidence!

---

http://xkcd.com/882/
PhD Student!
I think A may cause B, go investigate!
Yes professor!

I ran the analysis:
p > .05

But did you adjust for confounder Z?
Ehh, no
Let me get right back to you

After adjustment for Z, p < .05!
Yay! Let's publish a paper!
A solution?

Stop doing one study at a time!
What if we considered all outcomes?

Duloxetine vs. Sertraline for these 22 outcomes:

<table>
<thead>
<tr>
<th>Acute liver injury</th>
<th>Hypotension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myocardial infarction</td>
<td>Hypothyroidism</td>
</tr>
<tr>
<td>Alopecia</td>
<td>Insomnia</td>
</tr>
<tr>
<td>Constipation</td>
<td>Nausea</td>
</tr>
<tr>
<td>Decreased libido</td>
<td>Open-angle glaucoma</td>
</tr>
<tr>
<td>Delirium</td>
<td>Seizure</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>Stroke</td>
</tr>
<tr>
<td>Fracture</td>
<td>Suicide and suicidal ideation</td>
</tr>
<tr>
<td>Gastrointestinal hemorrhage</td>
<td>Tinnitus</td>
</tr>
<tr>
<td>Hyperprolactinemia</td>
<td>Ventricular arrhythmia and sudden cardiac death</td>
</tr>
<tr>
<td>Hyponatremia</td>
<td>Vertigo</td>
</tr>
</tbody>
</table>
All outcomes

Duloxetine vs. Sertraline

All these confidence intervals have been calibrated

All these confidence intervals have been corrected for multiple testing
All outcomes

Duloxetine vs. Sertraline

- Vertigo
- Vent. arr. & SCD
- Tinnitus
- Suicide & SI
- Stroke
- Seizure
- OA glaucoma
- Nausea
- Insomnia
- Hypothyroidism
- Hypotension
- Hyponatremia
- Hyperprolactinemia
- GI hemorrhage
- Fracture
- Diarrhea
- Delirium
- Decreased libido
- Constipation
- Alopecia
- ALI
- Acute MI
What if we consider all treatments?

<table>
<thead>
<tr>
<th>Type</th>
<th>Class</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Atypical</td>
<td>Bupropion</td>
</tr>
<tr>
<td>Drug</td>
<td>Atypical</td>
<td>Mirtazapine</td>
</tr>
<tr>
<td>Procedure</td>
<td>ECT</td>
<td>Electroconvulsive therapy</td>
</tr>
<tr>
<td>Procedure</td>
<td>Psychotherapy</td>
<td>Psychotherapy</td>
</tr>
<tr>
<td>Drug</td>
<td>SARI</td>
<td>Trazodone</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>Desvenlafaxine</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>duloxetine</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>venlafaxine</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>Citalopram</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>Escitalopram</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>Fluoxetine</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>Paroxetine</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>Sertraline</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>vilazodone</td>
</tr>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>Amitriptyline</td>
</tr>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>Doxepin</td>
</tr>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>Nortriptyline</td>
</tr>
</tbody>
</table>
Large-scale estimation for depression

- 17 treatments
- 17 * 16 = 272 comparisons
- 22 outcomes
- 272 * 22 = 5,984 effect size estimates
- 4 databases (Truven CCAE, Truven MDCD, Truven MDCR, Optum)
- 4 * 5,984 = 23,936 estimates
Propensity models for all comparisons (Truven CCAE)
Large-scale estimation for depression

Each estimate has full diagnostics and evaluation
Example 1

Fluoxetine vs. psychotherapy
Suicide ideation
Database: Truven MDCR

Calibrated HR = 1.05 (0.51 – 2.51)
Example 2

**Mirtazapine vs. Citalopram**

**Constipation**

Database: Truven MDCD

Calibrated HR = 0.90 (0.70 – 1.12)
Estimates are in line with expectations

11% of exposure-outcome pairs have calibrated $p < 0.05$

In literature, 85% have $p < 0.05$
Large-scale estimation for depression

• Each estimate produced with same rigor, and could be published as a paper
  – Propensity score adjustment
  – Cox regression
  – Calibrated using negative and positive controls
  – …

• Not “data-mining”!
  – Results should be interpreted considering multiple testing
  – This can’t be done for literature
OHDSI recommendations for evidence dissemination

✓ Address observation study bias
  Addressed by adjusting for confounding, and verifying bias was addressed. Disseminate your diagnostics and evaluations.

✓ Address publication bias
  Avoided by showing all tests that were performed, not just those with $p < 0.05$

✓ Address p-hacking
  Very hard to fine-tune analysis to one specific result
Population-level effect estimation

**Evidence Generation**
- Write and share protocol
- Open source study code
- Use validated software
- Replicate across databases

**Evidence Evaluation**
- Produce standard diagnostics
- Include negative controls
- Create positive controls
- Calibrate confidence interval and p-value

**Evidence Dissemination**
- Don’t provide only the effect estimate
- Also share protocol, study code, diagnostics and evaluation
- Produce evidence at scale
Building the LHC of observational research?
Join the journey