Observational research results in literature

• Individuals may produce good research studies
• In aggregate, the medical research system is a data-dredging machine
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
Repeat the Lee study in OHDSI

- Still had to infer many design features
Diagnose the propensity score distribution

Duloxetine vs. Sertraline

Results from Truven CCAE
Duloxetine: $n = 90,043$
Sertraline: $n = 175,950$
Diagnose covariate balance

Before stratification

Acer

After stratification

Age group 10-14

<table>
<thead>
<tr>
<th>Drug</th>
<th>Duloxetine</th>
<th>Sertraline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>0.2%</td>
<td>3.8%</td>
</tr>
<tr>
<td>After</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
P-value calibration

After calibration, 4% have $p < 0.05$ (was 16%)

Calibrated $p < 0.05$
Confidence interval calibration

Uncalibrated

Calibrated
Proposed evidence for stroke

Duloxetine vs. Sertraline

Results are comparable to Lee et al., but we provide the context to interpret the results.
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>
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
• $17 \times 16 = 272$ comparisons
• 22 outcomes
• $272 \times 22 = 5,984$ effect size estimates
• 4 databases (Truven CCAE, Truven MDCD, Truven MDCR, Optum)
• $4 \times 5,984 = 23,936$ estimates
Estimates are in line with expectations

11% of exposure-outcome pairs are significant once calibrated
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)
Propensity models for all comparisons (Truven CCAE, one outcome)
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
  – ...

• Calibration
  – Even if do not want to calibrate, must look at negative controls
Large-scale estimation for depression

• Not “data-dredging”!
  – Data-dredging is not about what you do but about what you *throw out*
  – This can’t be done for literature
  – Results should be interpreted considering multiple testing

• No reason not to carry out the other studies
  – Do not gain by not seeing them (blinding not relevant)
  – Studies are implicit in the data
Large-scale estimation for depression

• Bespoke studies
  – Wouldn’t it be best to optimize each study
  – Never get 10 or 100 parameters right
  – Still good to see the surface
    • Large-scale sensitivity analysis

• At the very least, publish every last parameter so it can be reproduced
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
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 that were significant

✓ Address CI-hacking
  Very hard to fine-tune analysis to one specific result
Join the journey

http://ohdsi.org