Reliable and reproducible effect size estimates at scale

Marc A. Suchard, M.D., Ph.D.

Departments of Biomathematics and Human Genetics
David Geffen School of Medicine at UCLA, and
Department of Biostatistics
UCLA Fielding School of Public Health
Popular antidepressants may raise stroke risk

Doctor, I'm starting on duloxetine, should I be worried about stroke?

Let me see what I find in the literature...
First hospitalization with ischemic stroke or intracranial hemorrhage in new users of SNRI vs. SSRI

- Taiwanese insurance claims
- 12 month washout
- Exclude dual-users, prior hx
- Hand-tailored propensity score stratification
- Time-varying, stratified Cox regression (ICD-9 433/434/436)

---

**Comparison of the Effects of Serotonin-Norepinephrine Reuptake Inhibitors Versus Selective Serotonin Reuptake Inhibitors on Cerebrovascular Events**

Yen-Chieh Lee, MD\textsuperscript{a,‡}; Chin-Hsien Lin, MD, PhD\textsuperscript{b,‡}; Min-Shung Lin, MD\textsuperscript{a}; Yun Lu, MSc\textsuperscript{c}; Chia-Hsuin Chang, MD, ScD\textsuperscript{c,d,‡}; and Jou-Wei Lin, MD, PhD\textsuperscript{e}

**ABSTRACT**

Background: Use of selective serotonin reuptake inhibitors (SSRIs) has been associated with an increased risk of intracranial hemorrhage. Depression is a highly prevalent neuropsychiatric disorder with a lifetime prevalence of more than 16% in the general population. Selective serotonin reuptake inhibitors (SSRIs) and serotonin-norepinephrine reuptake inhibitors (SNRIs) are often used to treat depression and anxiety disorders. However, there is limited evidence regarding the risks of intracranial hemorrhage associated with these medications.

Methods: This study used the Taiwan National Health Insurance Research Database to identify subjects with depression who initiated treatment with an SNRI or SSRI. The primary outcome was first hospitalization with ischemic stroke or intracranial hemorrhage. Propensity score stratification was used to adjust for baseline differences between the two groups. Time-varying, stratified Cox regression was used to estimate the adjusted hazard ratios.

Results: Among 76,920 patients who initiated SNRI treatment, 582,650 patients who initiated SSRI treatment, the adjusted hazard ratio for ischemic stroke was 1.01 (95% CI 0.90–1.12, p = 0.91).

Conclusion: The results suggest that SNRIs are not associated with an increased risk of ischemic stroke or intracranial hemorrhage compared to SSRIs.

---

**Crude Hazard Ratio (95% CI)**

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

- Do estimates generalize to different populations (databases)?
- How large is the residual bias?
- Do \( p \)-values have nominal characteristics?
- Do confidence intervals reflect “uncertainty” about the effect size?

Can I use this evidence to inform care for this patient?
Dulexotine (SNRI - 90,043) vs. sertraline (SSRI - 175,950) new-users

- US insurance claims (Truven CCAE)
- Dx of major depressive disorder and no bipolar or schizophrenia
- Propensity score via regularization on 58,285 potential covariates
- Fixed-time, stratified Cox model

Recommendation: always provide standard diagnostics

<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>
Calibrate the study evidence (under null)

Negative outcome controls (not caused by either treatment) help expose and control residual bias

<table>
<thead>
<tr>
<th></th>
<th>Crude - Uncalibrated</th>
<th>PS stratified - Uncalibrated</th>
<th>PS stratified - Calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>68% have $p &lt; 0.05$</td>
<td>16% have $p &lt; 0.05$</td>
<td>4% have $p &lt; 0.05$</td>
</tr>
</tbody>
</table>

$p$-value calibration models residual bias as exchangeable and adjusts for a (possibly) non-0 mean.
Trouble with **positive controls**:

- Few positive exemplars for a particular comparison
- Effect size never known certainty (and depends on population)
- We know they’re positive and change our behavior accordingly
Proposed positive control solution

- Start with a negative control (RR = 1)
Proposed positive control solution

- Inject simulated outcomes during exposure until desired RR is achieved
Proposed positive control solution

- Simulants behave like a ‘real’ event: **preserve confounding** by injecting outcomes for patients at high risk.
CI calibration models residual bias as exchangeable at a given RR and adjusts for a (possibly) non-0 mean at the estimated RR through a hierarchical model.
Results are comparable to Lee et al. across several populations (databases)

Enabled via open-source, quality-controlled tool standardization

But Lee et al. is one study among MANY in the literature . . .
85% have reported \( p < 0.05 \)

Also note unusual peak along boundary
What is going wrong? *p*-hacking

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!
Systematically evaluate many outcomes

- 22 adverse events with EU-ADR priority or depression-relevance
- All estimates calibrated and corrected for multiple testing
Systematically evaluate many treatments

<table>
<thead>
<tr>
<th>Type</th>
<th>Class</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>amitriptyline</td>
</tr>
<tr>
<td>Drug</td>
<td>Atypical</td>
<td>bupropion</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>citalopram</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>desvenlafaxine</td>
</tr>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>doxepin</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>duloxetine</td>
</tr>
<tr>
<td>Procedure</td>
<td>ECT</td>
<td>electroconvulsive therapy</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>Atypical</td>
<td>mirtazapine</td>
</tr>
<tr>
<td>Drug</td>
<td>TCA</td>
<td>nortriptyline</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>paroxetine</td>
</tr>
<tr>
<td>Procedure</td>
<td>Psychotherapy</td>
<td>psychotherapy</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>sertraline</td>
</tr>
<tr>
<td>Drug</td>
<td>SARI</td>
<td>trazodone</td>
</tr>
<tr>
<td>Drug</td>
<td>SNRI</td>
<td>venlafaxine</td>
</tr>
<tr>
<td>Drug</td>
<td>SSRI</td>
<td>vilazodone</td>
</tr>
</tbody>
</table>

- 17 treatments for depression with exposed $N \geq 10,000$
- Some comparable user-cohorts, some not (e.g. ECT)
Reliable and reproducible, large-scale estimation

\[
\binom{17}{2} \times 22 \text{ outcomes} \times 4 \text{ databases} = 11968 \text{ estimates}
\]

- 11\% of exposure-outcome pairs have \( p < 0.05 \)

And **no data dredging**! All estimates have **full diagnostics**, **evaluation** and **adjustment for multiple testing**.
Acknowledgments

Joint work with

- M. Schuemie and P. Ryan (J&J)
- P. Rijnbeek (Erasmus)
- D. Madigan and G. Hripcsak (Columbia)

Financial support

- Alfred P. Sloan Fellowship
- John Simon Guggenheim Fellowship
- NSF IIS 1251151 (BigData) and DMS 1264153
- NIH R01 AI107034 and R01 HG006139

All protocols and tools at http://www.github.com/OHDSI