# Systematic Reviews:

# How Accurate Are They and Can We Do Better?

Authors:
Gerald Borok, PhD, MPH
Todd Feinman, MD
Iris Tam, PharmD

Gerald M. Borok, PhD, MPH
Senior Director, Client Solutions
Doctor Evidence, LLC
gborok@doctorevidence.com
growthevidence.com

Presented at 11th G-I-N (Guideline International Network) Annual Conference
Melbourne, Australia
August 20-23, 2014

# Disclosure of Interests

- Gerald Borok, PhD, MPH
- Sr. Director, Client Solutions
- Employed by Doctor Evidence, LLC, a specialty evidence-based medicine software platform and company with a mission to provide stakeholders across the healthcare ecosystem with the most timely and accurate relevant medical evidence and related analytics.
- Projects funded by range of healthcare clients including non-profit healthcare organizations, pharmaceutical and biotechnology industry. Funding for this systematic review validation project was by a pharmaceutical client as part of a larger scope of work.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Presentation Outline

- Background
- Objectives
- Methods
- Results
- Discussion
- Implications for Guideline Developers

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Background

- The Institute of Medicine (IOM) states that trustworthy guidelines depend on systematic reviews (SRs) of published evidence.[1] However, IOM reports that SRs, if used at all, for guideline development have variable quality and transparency. [2]
- Different methodologies and approaches used in the development of SRs can affect the accuracy of results, which in turn can impact conclusions and recommendations.
- The body of literature evaluating the quality of, and alternative approaches to conducting systematic reviews is limited.
- On the issue of data extraction of quantitative and other critical data for SRs, IOM [2] and Cochrane [3] recommend data extraction independently by at least two persons. IOM also recommends a fair procedure for resolving discrepancies.


1 IOM (Institute of Medicine). 2011. Clinical Practice Guidelines We Can Trust. Washington, DC: The National Academies Press.

2 IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: The National Academies Press.

3 Higgins JPT, Deeks JJ (editors). Chapter 7: Selecting studies and collecting data. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.


GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION

GROWTH

# Background

- Data extraction is a relatively under-researched area of the systematic review process compared to information retrieval, the assessment of bias, and methods of synthesis. [1]
- Research on extraction error rates has found the following:
  - Jones et al. [2] reported a high rate (20 of 34 SRs) of data extraction errors. The authors used an experienced statistician to retrospectively repeat the data extraction in SRs.
  - Horton et al. [3] reported data extraction error rates of 28.3% to 31.2%, and were similar across levels of reviewer experience.
  - Goetzsche et al. [4] evaluated 27 meta-analyses using standard mean differences, and reported they could not replicate the results for at least 1 of the 2 trials (randomly selected from each meta-analysis article) in 37% of the meta-analyses. Extraction problems were erroneous number of patients, means and standard deviations.

[1] Carroll C et al. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Research Notes 2013;6:539.

[2] Jones AP et al. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol 2005;58:741-742.

[3] Horton J et al. Systematic review data extraction cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol 2010;63(3):289-298.

[4] Goetzsche PC et al. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298(4):430-437.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Background

- Limited research on different extraction approaches have reported the following results:
  - Carroll et al. [1] found error rates between 8% and 17% across three binary outcomes between three reviews of the same studies.
    - They recommended SRs use double-data extraction.
  - Bucemi et al. [2] reported that independent data extraction by two analysts of the same article resulted in fewer errors than data extraction by a single analyst followed by verification by a second person (relative difference: 21.7%, p=.019).
  - Heywood et al. [3] reported strong agreement in data extraction for the same set of selected articles between three reviewers with different expertise using an electronic database with careful instruction and training.
    - They also identified the benefit of the electronic database to facilitate data extraction and data entry in one step.

[1] Carroll C et al. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Research Notes 2013;6:539.

[2] Bucemi N et al. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol 2006;59:67-703.

[3] Heywood KL et al. Reviewing measures of outcome: reliability of data extraction. J Eval Clin Pract 2004;10:329-337.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Background

- Research on sources of errors, including database and analytic issues:
  - Goetzsche et al.[1] reported for the 10 meta-analyses with important errors, they could not replicate the authors' pooled result in 70% of the meta-analyses.
  - Jones et al.[2] reported that in 34 SRs:
    - Incorrect calculations made when converting data in primary articles into data required for the review occurred in 2 SRs
    - Misinterpretation of data that was reported in the primary article occurred in 7 SRs
    - Data entry errors in 3 SRs
    - Data reported in SR differed from the published article in 1 SR
    - Results from one trial given as results from another trial in 2 SRs

[1] Goetzsche PC et al. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298(4):430-437.

[2] Jones AP et al. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol 2005;58:741-742.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Objectives

- To evaluate a digital evidence-based technology solution for conducting systematic reviews (SRs) and meta-analyses (MAs) versus traditional methods.
- Specifically, we sought to determine the accuracy of data extraction and analytical capability of the digital platform.

# Methods

- This research is one of three SRs selected as part of validation projects to compare published SRs vs. the Doctor Evidence (DRE) technology-based platform in different therapeutic areas:
  - Rheumatoid arthritis (**the focus for this presentation**)
  - Lung cancer screening
  - COPD
- The published SR using traditional literature search and data extraction methods was compared to the DRE technology-supported processes to determine extraction and analytic accuracy. Specifically:
  - Data was extracted from articles cited in the selected SR, using a dual, blinded extraction plus quality control review process supported by the DRE IT platform.
  - Conducted series of 120 MAs to replicate direct and indirect MAs in published SR/MA, using the DRE integrated analytic platform.
- Data extraction and analyses done in the platform were compared to the published results. Discrepancies were evaluated for causes.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Methods

- SR selected for validation project because included by client as part of literature review. SR selection criteria included:
  - Comprehensive set of RCTs included in SR
  - Inclusion of full array of biologics
  - Thorough description of methodology
  - Recently published
- Citation: Not provided at this time because of the preliminary nature of the findings – have not yet contacted SR author for review and comment.
- Funding source for SR: Not the same source as the client requesting the validation project.
- Definition of outcome: American College of Rheumatology (ACR) criteria scores 20, 50 and 70. Defined as:
  - A 20/50/70 percent improvement in tender or swollen joint counts, and
  - A 20/50/70 percent improvement in three of the following five measures: patient and physician global assessments, pain, disability, and an acute phase reactant (sedimentation rate).

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Results

- Overall, a large number of errors were found in the published SR/MA:
  - 71 of 120 MAs (59.2%).
- For the 71 errors, there were four specific types of errors attributable to SR:
  - Incorrect population denominator, affecting input values for meta-analyses (43.7%)
  - Mislabeling a study - mismatch between reference in analytic section and citation in reference list - and therefore cannot be included in DRE replication analysis (16.9%)
  - Mismatch between results reported by SR and results in original studies – causing input values for SR meta-analyses to be erroneous (16.9%)
  - Miscalculation of odds ratios due to errors entering data into statistical software (when SR input values matched original studies and DRE) (22.5%)
- After correcting errors and re-running meta-analyses, two treatment comparison findings which were nonsignificant in the SR would have been statistically significant.
  - Rituximab combination therapy was significantly better than DMARDs alone for ACR 70.
  - Etanercept was significantly better than Adalimumab for ACR 20.
  - Potential impact on article Abstract: would have changed one of the four results statements (non-significant to significant), and added a fifth results statement.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Results Of Confirmatory Odds Ratio Calculator Analysis

- There was a substantial (nearly two-fold) difference observed for an ACR 70 result in direct meta-analysis of monotherapy vs. placebo) between SR and DRE, but input values were identical for DRE meta-analysis and SR article.
- This large observed difference required a validation of results generated by the DRE software.
- Two external, web-based Odds Ratio (OR) Calculators were used to confirm the accuracy and validity of the DRE software.
- Both calculators produced the exact same results (OR of 21.860 for the ACR 70 example) as the DRE software, confirming its validity.
- In contrast, the result reported by SR (OR of 40.714) for the same ACR 70 comparison was not confirmed.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Discussion

- Published SR used a single extractor + second reviewer process vs. DRE dual extractor + QC reviewer process supported by digital technology.
- Published SR used a separate analytic software package vs. DRE analytic package which is part of an integrated extraction-analytic database platform.
- These differences in the extraction and analytic components of the SR/MA contributed to (1) the observed errors in the published SR, and (2) the ability of the DRE technology to identify SR errors.
- Inaccuracies in SRs/MAs may lead to erroneous safety/efficacy conclusions in published reviews and subsequent guidelines.
- Use of digital technologies and data extraction with high quality control, coupled with an analytic package in an integrated platform, may be more effective and accurate for conducting SRs/MAs.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Implications for Guideline Developers

- Guideline developers need to be aware of the source and types of errors in SRs identified in this evaluation, and reconsider reliance on SRs which do not use the rigorous data extraction and analytic processes recommended by IOM and Cochrane.
- Our results indicate that the use of a digital evidence-based platform can significantly improve accuracy and quality of SRs/MAs.
- Thus, guideline developers may consider adopting this technology for developing trustworthy guidelines to meet IOM standards.

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Acknowledgements

Gerald Borok, PhD, MPH
Sr. Director, Client Solutions
Doctor Evidence
gborok@doctorevidence.com

Iris Tam, PharmD
Director, Managed Care
Medical Communications
Genentech
tam.iris@gene.com

Todd Feinman, MD
CMO & Founder
Doctor Evidence
tf@doctorevidence.com

www.doctorevidence.com
http://growthevidence.com/

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Additional Information

# Results Section of Abstract in SR publication

- DRE identified two treatment comparison findings which were nonsignificant in the SR, but would have been statistically significant if analyzed correctly.
  - Rituximab combination therapy was significantly better than DMARDs alone for ACR 70 – highlighted by SR as an "exception" in Abstract.
  - Etanercept (monotherapy) was significantly better than Adalimumab (monotherapy) for ACR 20 – not included in Abstract because not statistically significant.
- **Results section of Abstract:** "The systematic review identified 10,625 citations, and after a review of 2450 full-text papers, there were 29 and 14 eligible studies for the combination and monotherapy meta-analyses, respectively. In the combination analysis, all licensed bDMARD combinations had significantly higher odds of ACR 20/50/70 compared to DMARDs alone, **except for the rituximab comparison, which did not reach significance for the ACR 70 outcome (based on the 95% credible interval).** The etanercept combination was significantly better than the tumor necrosis factor-α inhibitors adalimumab and infliximab in improving ACR 20/50/70 outcomes, with no significant differences between the etanercept combination and certolizumab pegol or tocilizumab. Licensed-dose etanercept, adalimumab, and tocilizumab monotherapy were significantly better than placebo in improving ACR 20/50/70 outcomes. Sensitivity analysis indicated that including studies outside the target population could affect the results."

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION GROWTH

# Summary of Error Rate by Reason and Attribution: SR vs. DRE Meta-analyses Comparison

| SR vs. DRE Comparison Reasons for Difference | Attribution of Errors | Total Number of Meta-analyses with Differences | Number SR Incorrect | Percent SR Incorrect |
|---|---|---|---|---|
| Incorrect population denominator (affecting accuracy of meta-analyses). | SR incorrect | 71 | 31 | 43.7% |
| Mislabeled a study (different RCT in results section vs. citation in reference section). Study therefore not available for inclusion in DRE replication meta-analysis. | SR incorrect | 71 | 12 | 16.9% |
| Results for one study were incorrect vs. original article, affecting accuracy of input values for meta-analyses. | SR incorrect | 71 | 12 | 16.9% |
| Miscalculation of odds ratios in SR meta-analyses (when input values the same for SR, original studies and DRE). | SR incorrect | 71 | 16 | 22.5% |
| All Reasons for Differences | SR incorrect | 71 | 71 | 100.0% |

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Summary of Error Rate by Type of Meta-analysis and Attribution for Tables S1, S5, S2, S6

| SR Table | Analysis | Attribution of Errors | Total Number of Meta-analyses | Number SR Incorrect | Percent SR Incorrect |
|---|---|---|---|---|---|
| **S1** | Direct meta-analysis of ACR 20, 50 and 70 outcomes: combination therapy. | SR incorrect | 48 | 31 | 64.5% |
| **S5** | Direct meta-analysis of ACR 20, 50 and 70 outcomes: licensed DMARD monotherapy vs. placebo in DMARD-experienced patients | SR incorrect | 18 | 10 | 55.5% |
| **S2** | Bucher indirect meta-analysis of ACR 20, 50 and 70 outcomes: combination therapy | SR incorrect | 42 | 24 | 57.1% |
| **S6** | Bucher indirect meta-analysis of ACR 20, 50 and 70 outcomes: licensed DMARD monotherapy vs. placebo in DMARD-experienced patients | SR incorrect | 12 | 6 | 50.0% |
| **TOTAL** | All Meta-analyses | SR incorrect | 120 | 71 | 59.2% |

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH

# Example of Results for Replication of Table S5 ACR 70

| Treatment | SR Fixed Effect OR* v PLA** (95% CI) | DRE Fixed Effect OR v PLA (95% CI) | SR Random Effect OR v PLA (95% CI) | DRE Random Effect OR v PLA (95% CI) | Explanatory Notes |
|---|---|---|---|---|---|
| **ADA 40 mg/2 weeks monotherapy vs. Placebo** | *10.861 (3.045 to 38.736)* | 9.04 (2.68 to 30.49) | *10.126 (2.837 to 36.145)* | 8.88 (2.63 to 30.03) | Van de Putte (2004): Discrepancy between SR and DRE results, when input values are the same for SR and DRE (as per Table 10 in SR). Possible miscalculation by SR. |
| **ETN 2 x 25 mg/week monotherapy vs. Placebo** | *25.714 (3.215 to 205.639)* | 14.36 (1.82 to 113.38) | *25.714 (3.215 to 205.639)* | 14.36 (1.82 to 113.38) | Moreland (1999) Large discrepancy between SR and DRE results, when input values are the same for SR and DRE (as per Table 10 in SR). Possible miscalculation by SR. |
| **TOC 8 mg/kg/4 weeks monotherapy** | *40.714 (2.276 to 728.176)* | 28.07 (9.69 to 81.32) | *40.714 (2.276 to 728.176)* | 28.07 (9.69 to 81.32) | Nishimoto (2004) Large discrepancy between SR and DRE results, when input values are the same for SR and DRE (as per Table 10 in SR). Possible miscalculation by SR. |

*OR: Odds Ratio **PLA: Placebo

GUIDELINE & RESI for TRANSPARENC GR

# Example of Results for Replication of Table S6 ACR 20

| Treatment | Control | SR Fixed Effect OR* v DMARD (95% CI) | DRE Fixed Effect OR v DMARD (95% CI) | SR Random Effect OR v DMARD (95% CI) | DRE Random Effect OR v DMARD (95% CI) | Explanatory Notes |
|---|---|---|---|---|---|---|
| **ETN 2x25 mg/week** | **ADA 40 mg/2 weeks** | ***2.148 (0.818 to 5.639)*** | **2.769 (1.072 to 7.155)** | ***2.140 (0.816 to 5.616)*** | **2.777 (1.074 - 7.180)** | Discrepancy due to difference between SR and DRE for ADA vs. PLA result (Van de Putte, 2004) when input values were the same for SR and DRE (per table 10 in SR). ETN vs. PLA results the same for SR and DRE. **This difference resulted in difference in meta-analysis results: DRE results for Fixed and Random Effect ORs are statistically significant; SR result not statistically significant.** |
| **ETN 2x25 mg/week** | **TOC 8 mg/kg/4 weeks** | *0.404 (0.105 to 1.555)* | 0.404 (0.105 to 1.555) | *0.404 (0.105 to 1.555)* | 0.404 (0.105 - 1.555) | Results the same for SR and DRE analyses. |

*OR: Odds Ratio

GUIDELINE & RESEARCH ORGANIZATIONS WORLDWIDE for TRANSPARENCY & HARMONIZATION
GROWTH