ebhc CONFERENCE 2026

THE ECOSYSTEM OF EVIDENCE

Navigating the future
of the ecosystem of evidence

Artificial intelligence and digital transformation
in research and healthcare

11th EBHC International Conference
Taormina, 21st – 24th October 2026

#EBHC2026

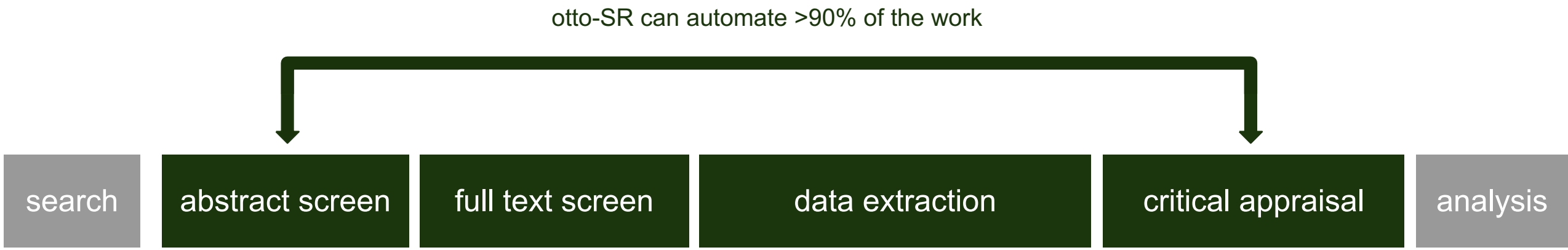
30th GIMBE 1996-2026 EVIDENCE FOR HEALTH

International Society for Evidence-Based Health Care

Automation of systematic reviews with large language models

Christian Cao
University of Toronto
Otto Labs Inc.

Background



SRs are incredibly cost- and time-intensive

- Often take over a year to complete (mean 67.3 weeks)
- Average cost: $141,194.80 USD per report. Some cases span upwards of $1.4M

Countless automation tools, but none can replace human reviewers… until now

Aims

1) Validate the performance of otto-SR in title/abstract and full-text screening

1) Validate the performance of otto-SR in data extraction

1) Validate the performance of otto-SR in reproducing and updating the primary analysis of an entire issue of Cochrane reviews

Methods (Phase 1-2)

Overview

Comparative diagnostic test accuracy study using three different arms:

- otto-SR (leveraging GPT4.1)
- Dual human reviewers (*calibrated*)
  - Graduate and postgraduate level
- Elicit, DistillerSR, NestedKnowledge (commercially available software)

Reference Standard

- Screening: Article inclusion decisions after full-text screening by the original review authors
- Extraction: Values from the original review, with additional correction/adjudication

Datasets

Datasets encompassed reviews of prevalence, intervention benefits, prognosis, and diagnostic test accuracy; covering eight unique clinical domains

Screening

- Otto-SR - evaluated on **32,357 citations**
- Dual humans, Software - evaluated on a powered sample of **1,767 citations**

Extraction

- Otto-SR - evaluated on **4,559 data points, 495 studies**
- Dual humans, Software - evaluated on a powered sample of **1,453 data points, 156 studies**

Results: Phase 1 Screening

Abstract

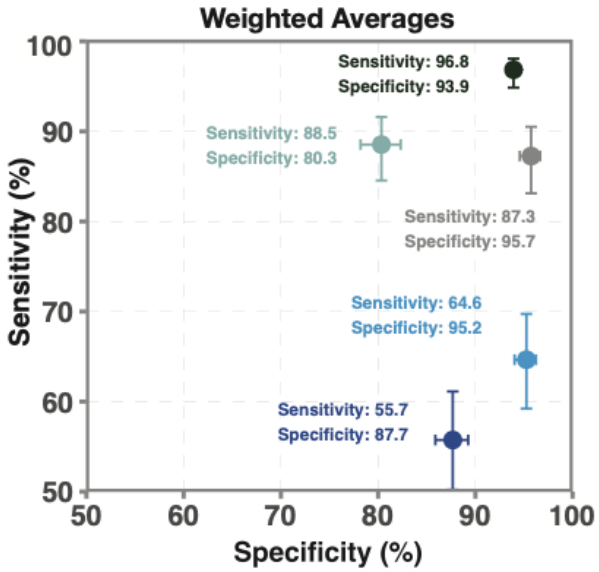
Otto-SR
- **96.7% sensitivity**
- 93.9% specificity

Dual human
- 87.3% sensitivity
- **95.7% specificity**

Elicit
- 88.5% sensitivity
- 84.2% specificity

DistillerSR
- 61.4% sensitivity
- 95.3% specificity

Nested Knowledge
- 55.7% sensitivity
- 87.7% specificity



Full-text

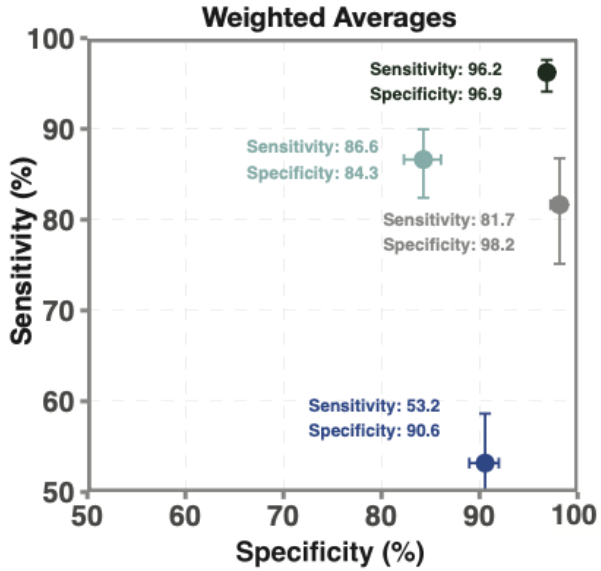
Otto-SR
- **96.2% sensitivity**
- 96.9% specificity

Dual human
- 81.7% sensitivity
- **98.1% specificity**

Elicit
- 86.6% sensitivity
- 84.3% specificity

DistillerSR
- 81.7% sensitivity
- 98.2% specificity

Nested Knowledge
- 53.2% sensitivity
- 90.6% specificity



Results: Phase 2 Extraction

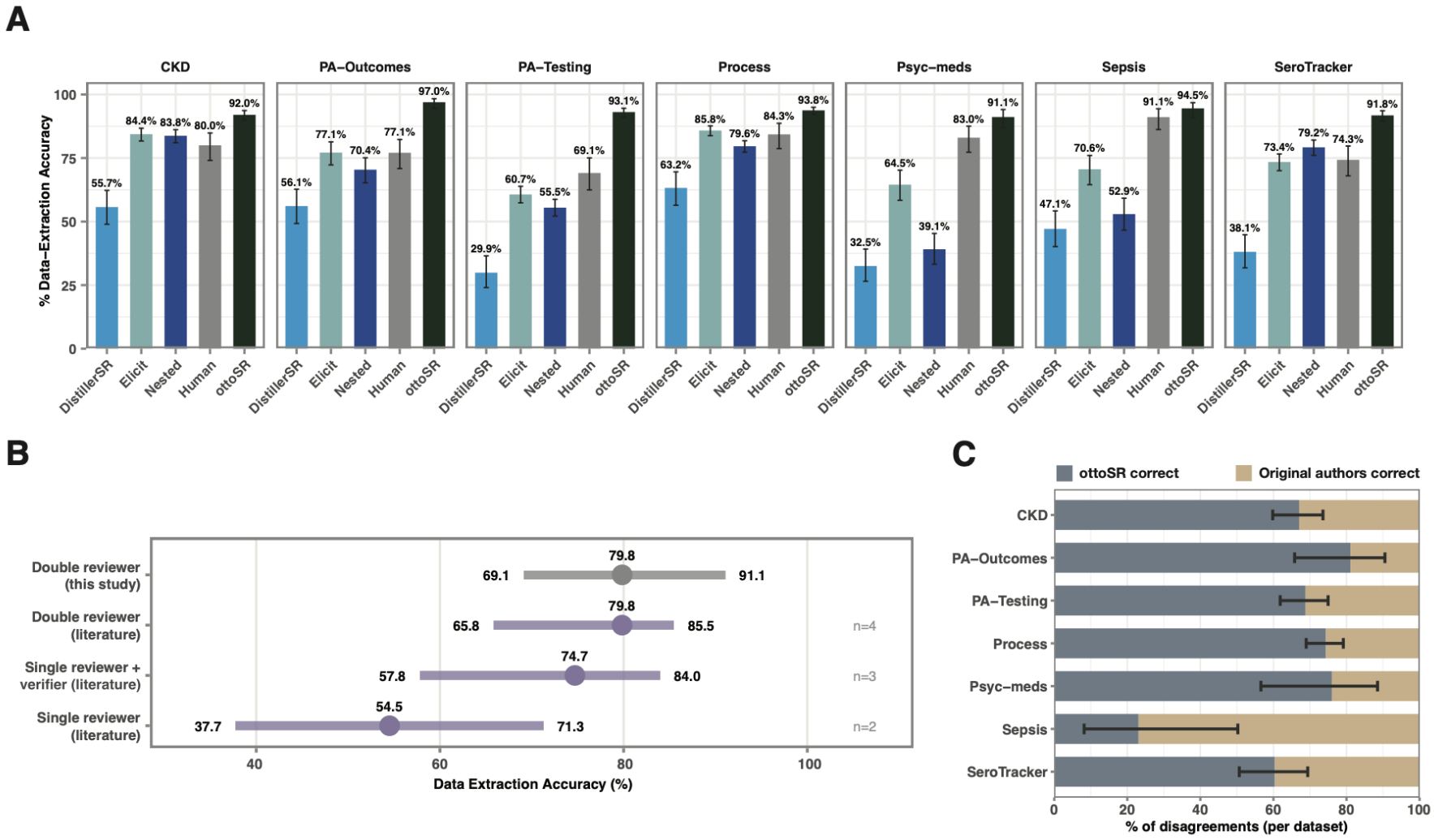
Human: 79.7%

otto-SR: 93.1%

Elicit: 72.0%

Nested Knowledge: 71.2%

DistillerSR: 46.1%



Methods: Phase 3 Reproducibility

Overview

Wanted to assess real-world impact of our work

Reproduced the entire April 2024 Cochrane issue (n=14 reviews), from initial search to meta-analysis

Specifically focused on primary outcome to reduce inclusion ambiguity (*i.e., avoid dealing with discovering new outcomes to analyze)*

Determined 12 reviews eligible for further analysis, comprised **146,276 citations.**

Completed in **under 2 days.**

Analytical Overview

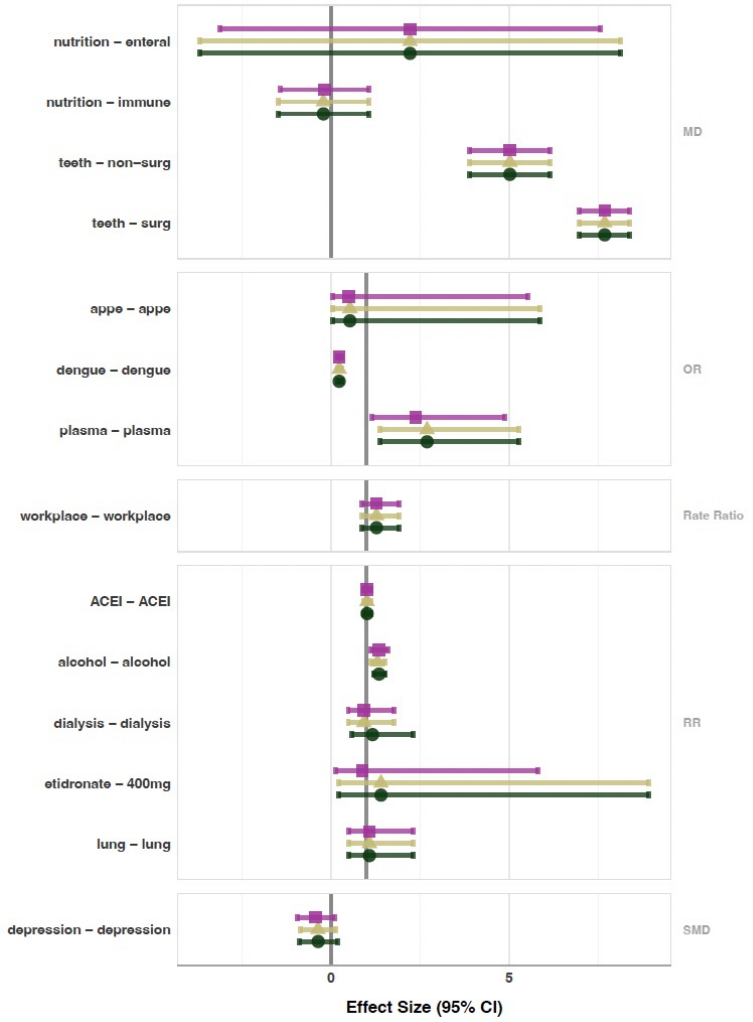
Cochrane:

- Same studies and raw data provided by Cochrane authors
- Excluded studies not captured by database search + implicit data requests

Compared final meta-analyzed results under three scenarios:

- **Matched:** Applied the same date cut-offs, only meta-analyzed the studies included in the original Cochrane review
- **Expanded:** Applied same date cut-offs, meta-analyzed all studies included by otto-SR
- **Update:** Updated reviews to April 2025

Results: Phase 3 Reproducibility - Matched



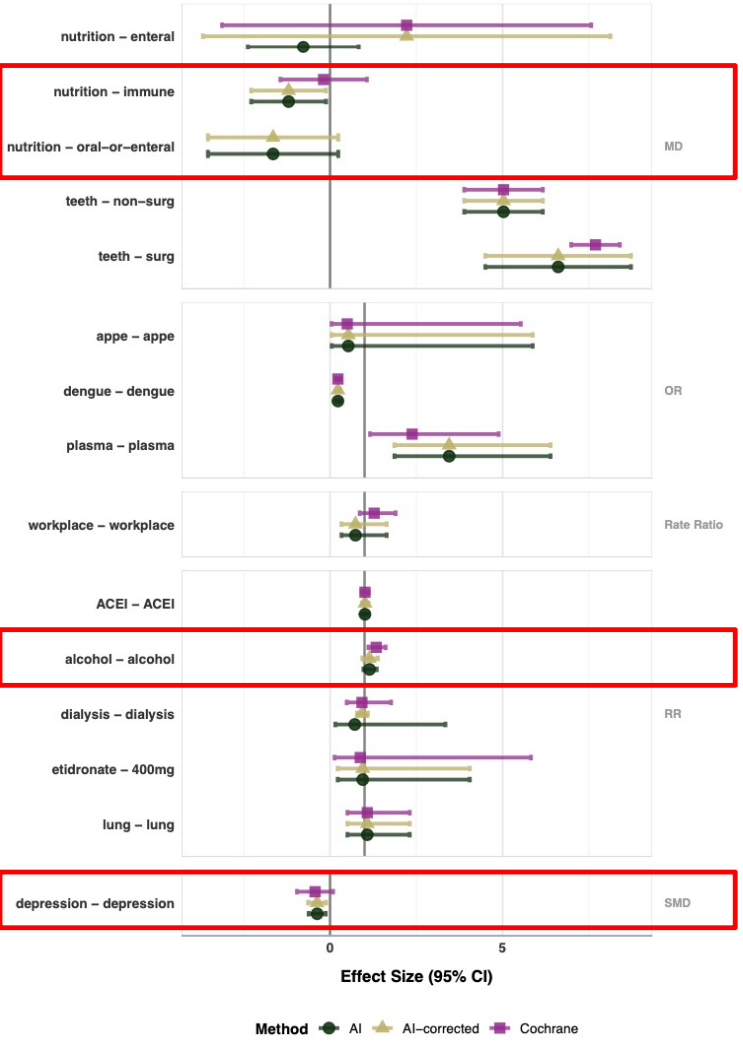
All Cochrane, Otto-SR, and 'Corrected' (gold-standard) had similar meta-analyzed effect values + overlapping 95% CIs

Otto-SR Screening Agent: 64/64 included studies (100% sensitivity; included all Cochrane studies)

Otto-SR Extraction Agent: 60/64 (93.8% sensitivity)

- 2 studies - outcome data reported in supplementary tables (inaccessible)
- 2 studies (1 article) - *otto-SR* failed to extract outcome data (false omission)

Results: Phase 3 Reproducibility - Expanded

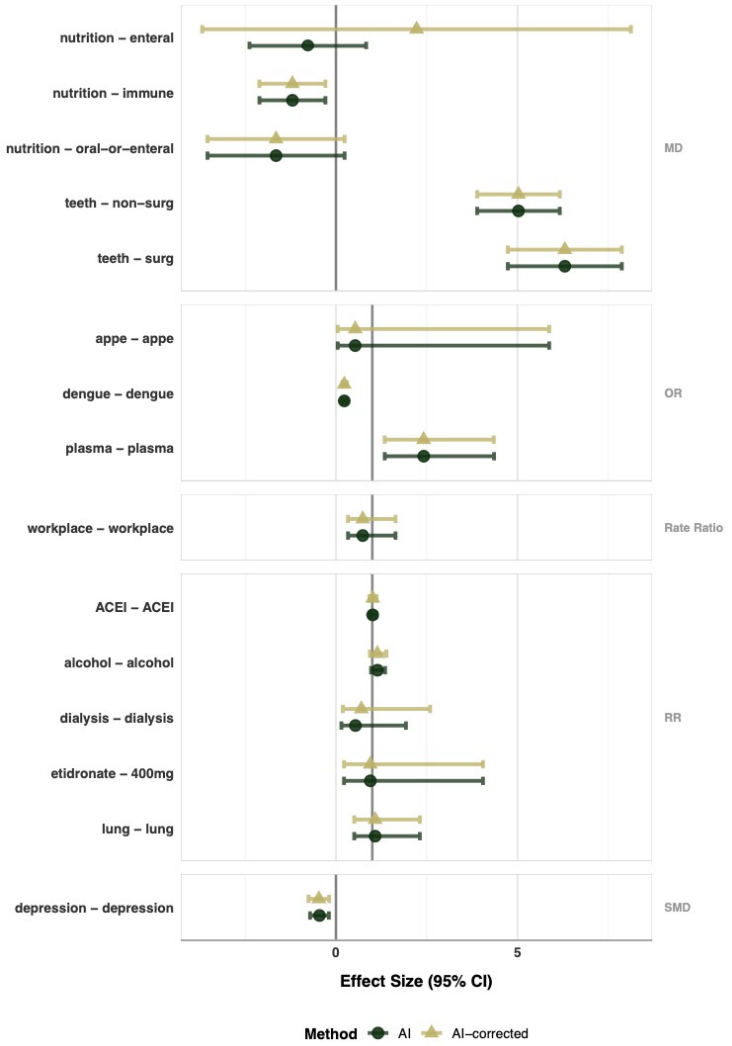


Included an additional 54 eligible studies that were likely missed by the original authors across the 12 reviews (median 2, IQR 1 to 6.5)

Key findings

1. New statistical significance in 2 studies
2. 1 study lost statistical insignificance
3. Identified a new analytical comparison

Results: Phase 3 Reproducibility - Update



Included an additional 68 eligible studies that were likely missed by the original authors across the 12 reviews (median 2.5, IQR 1 to 8)

Same key findings as matched

1. New statistical significance in 2 studies
2. 1 study lost statistical insignificance
3. Identified a new analytical comparison

Separately, 9 false positive studies were identified

- Of these, 7 (77.7%) may have contained relevant data accessible through author correspondence.

Limits

- Phase 3 reproducibility experiment limited to primary outcome (for feasibility)

- otto-SR was limited to the main text; it did not have built-in supplementary material retrieval capabilities during time of evaluation
  - *We penalized otto-SR's data extraction performance in instances where supplementary materials were needed*

- Applied a validated LLM-as-a-judge system (which may still introduce some errors)

- Potential concerns around data contamination - retrospective reviews may have been included in the model pretraining corpora
  - *Datasets used in screening/data extraction validation may be subject to contamination*
  - *The Phase 3 Cochrane Reproducibility used reviews published in April 2024, which were closed access during the time of GPT4.1 model pretraining*

Conclusions

otto-SR can be used to:

Create true “living” reviews.

- Updates happen the same day, instead of requiring months-years

Rapidly conduct new reviews*

- Rapid review methodology: 1 reviewer vote
- More prospective validation needed to assess performance in traditional systematic reviews

Future Research

Data availability & reproducibility

- Most SRs are not reproducible (including Cochrane!)
- Better data sharing standards? *Although this can create its own issues*

Further review component automation

- Risk of bias / critical appraisal
- Meta-analysis & meta-regression