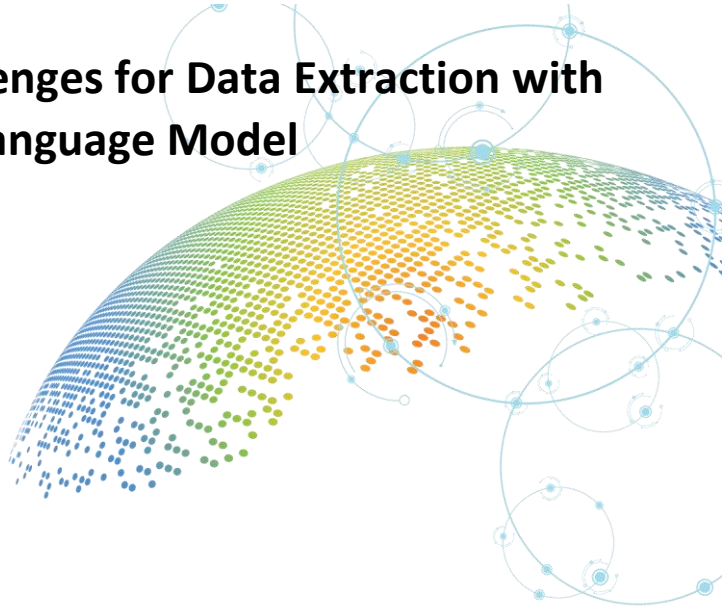
Opportunities and Challenges for Data Extraction with a Large Language Model

Gerald Gartlehner

October 17, 2024



1

Declaration of Conflict of Interest

I have no actual or potential conflict of interest in relation to this presentation.

The research was funded by the RTI International Innovation Office and the US Agency for Healthcare Research and Quality (AHRQ).

2

Data Extraction

- The process of transcribing data from primary studies into standardized tables.
- Conducted by **two investigators** independently or through extraction by one person and verification by another.
- It **varies in complexity** from copying and pasting to transformations or calculations to obtain data.
- Data extraction is **time-consuming, costly, tedious, and error-prone**.
- Up to 63% of studies included in systematic reviews have at least one data extraction error.

Mathes T, Klassen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152.

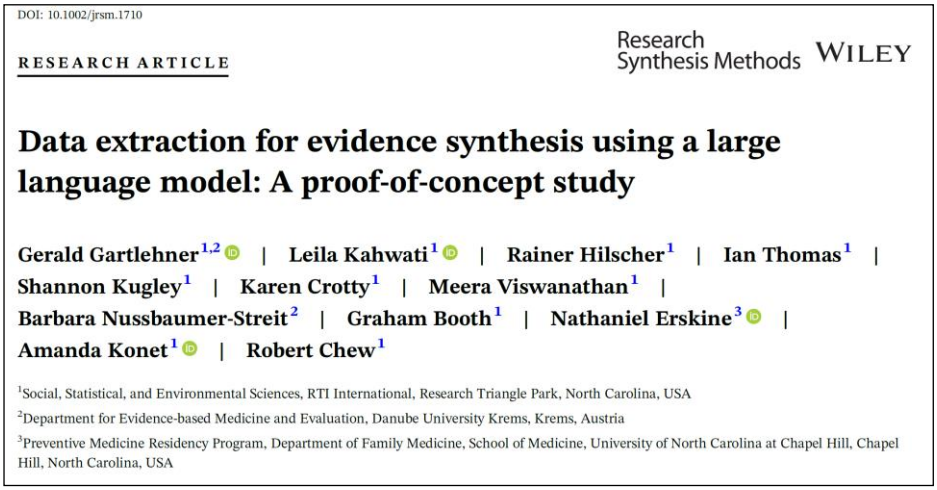
3

Use of AI for Data Extraction

- Previous methods have mostly focused on **natural language processing** using statistical models (support vector machine, Bayesian) or deep neural networks.
- Tools **require large labeled training** sets for the machine to “learn” and often have **not achieved** sufficient accuracy.
- **Large Language Models** allow **zero-shot applications** for data extraction: no training or programming is necessary.

4

Test-of-Concept



DOI: 10.1002/jrsm.1710

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Data extraction for evidence synthesis using a large language model: A proof-of-concept study

Gerald Gartlehner[1,2] | Leila Kahwati[1] | Rainer Hilscher[1] | Ian Thomas[1] | Shannon Kugley[1] | Karen Crotty[1] | Meera Viswanathan[1] | Barbara Nussbaumer-Streit[2] | Graham Booth[1] | Nathaniel Erskine[3] | Amanda Konet[1] | Robert Chew[1]

[1]Social, Statistical, and Environmental Sciences, RTI International, Research Triangle Park, North Carolina, USA

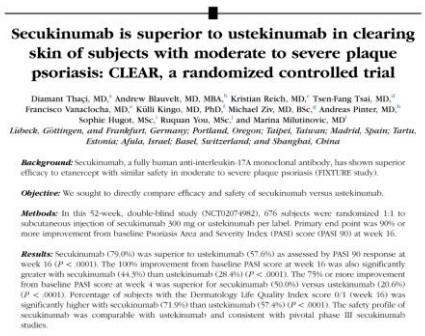
[2]Department for Evidence-based Medicine and Evaluation, Danube University Krems, Krems, Austria

[3]Preventive Medicine Residency Program, Department of Family Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Study Design

- **Validation study** to compare the performance of LLM for data extraction against a reference standard
- Reference standard: **Enhanced manual data extraction by humans**
- Convenience sample of **10 open-access journal publications** of RCTs from a previous review provided as PDFs
- **16 data elements** including study and population characteristics, outcomes data, participant flow, etc.
- **Outcomes:** Accuracy of data extracted by LLM, reliability, and types of errors

Data Sources



Secukinumab is superior to ustekinumab in clearing skin of subjects with moderate to severe plaque psoriasis: CLEAR, a randomized controlled trial

Diamant Thaçi, MD,[a] Andrew Blauvelt, MD, MBA,[b] Kristian Reich, MD,[c] Tsen-Fang Tsai, MD,[d] Francisco Vanaclocha, MD,[e] Killi Kingo, MD, PhD,[f] Michael Ziv, MD, BSc,[g] Andreas Pinter, MD,[h] Sophie Hugot, MSc,[i] Ruquan You, MSc,[i] and Marina Milutinovic, MD[i]

Lübeck, Göttingen, and Frankfurt, Germany; Portland, Oregon; Taipei, Taiwan; Madrid, Spain; Tartu, Estonia; Afula, Israel; Basel, Switzerland; and Shanghai, China

Background: Secukinumab, a fully human anti-interleukin-17A monoclonal antibody, has shown superior efficacy to etanercept with similar safety in moderate to severe plaque psoriasis (FIXTURE study).

Objective: We sought to directly compare efficacy and safety of secukinumab versus ustekinumab.

Methods: In this 52-week, double-blind study (NCT02074982), 676 subjects were randomized 1:1 to subcutaneous injection of secukinumab 300 mg or ustekinumab per label. Primary end point was 90% or more improvement from baseline Psoriasis Area and Severity Index (PASI 90) at week 16.

Results: Secukinumab (79.0%) was superior to ustekinumab (57.6%) as assessed by PASI 90 response at week 16 ($P < .0001$). The 100% improvement from baseline PASI score at week 16 was also significantly greater with secukinumab (44.5%) than ustekinumab (28.4%) ($P < .0001$). The 75% or more improvement from baseline PASI score at week 4 was superior for secukinumab (50.0%) versus ustekinumab (20.6%) ($P < .0001$). Percentage of subjects with the Dermatology Life Quality Index score 0/1 (week 16) was significantly higher with secukinumab (71.9%) than ustekinumab (57.4%) ($P < .0001$). The safety profile of secukinumab was comparable with ustekinumab and consistent with pivotal phase III secukinumab studies.

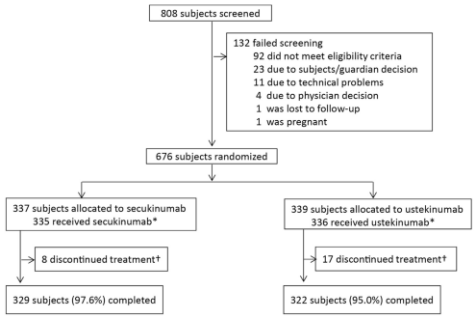


Table I. Baseline demographic and clinical characteristics

Characteristic	Secukinumab 300 mg (n = 337)	Ustekinumab (n = 339)
Age, y	45.2 ± 13.96	44.6 ± 13.67
Male gender	229 (68.0)	252 (74.3)
Race		
Caucasian	299 (88.7)	288 (85.0)
Other	38 (11.3)	51 (15.0)
Weight, kg	87.4 ± 19.95	87.2 ± 22.11
BMI, kg/m^2	29.1 ± 5.87	29.0 ± 6.69
Time since psoriasis diagnosis, y	19.6 ± 12.90	16.1 ± 11.24
PASI score	21.7 ± 8.50	21.5 ± 8.07
Body surface area involved, %	32.6 ± 17.78	32.0 ± 16.80
IGA mod 2011 score		
4 (Severe disease)*	130 (38.6)	125 (36.9)
Psoriatic arthritis reported	69 (20.5)	54 (15.9)
Previous systemic psoriasis treatment		
Any	225 (66.8)	231 (68.1)
Conventional agent†	218 (64.7)	223 (65.8)
Biologic agent	48 (14.2)	44 (13.0)
Failed biologic agent	36 (10.7)	34 (10.0)

7

7

Prompt Engineering

- During a pilot phase, we developed clear definitions for each of the 16 data elements.
- Iterative engineering of prompts based on definitions of data elements.

Variable „First author“: The last name of the first author

Prompt: The last name of the first author, styled as a proper noun with first letter capitalized

8

8

Accuracy

- For **160 data elements**, data were available in sample publications on **157**.
- When **data were available**, Claude successfully extracted the pertinent information with **96.2% accuracy** (151 out of 157 cases).
- When **data were lacking**, Claude accurately reported the absence in **100%** of the instances (3 out of 3 cases).
- The overall **accuracy was 96.3%.**
- In several cases, Claude detected minor **errors of humans**.

Types of Errors

1 major error

Incorrect (different dosing) and made up ("hallucinated") data for treatment group

1 minor error

Rounding error of standard deviation

Missed data

In 4 instances

Human Errors

Out of 160 data elements in the reference standard, Claude found 21 minor errors in in human data extractions

Mean duration of disease Ixekizumab: 18.0 **(1.1)**
Mean duration of disease Ixekizumab: 18.0 **(11.1)**

N (%) Female Placebo: 23 **(39.6)**
N (%) Female Placebo: 23 **(60.4)**

Test-Retest Reliability

- 4 weeks after first data extraction, we reran the same prompts for the same journal publications.
- **Proportions of errors** were similar: **3.7% vs. 3.1%.**
- **Agreement** between test and retest**: 93.4%**
- **But:** errors occurred for **different variables** during the replication, except in 1 instance.

Limitations

- Conducted in a **controlled environment** with involving data scientists.
- Focus on **RCTs of pharmacologic interventions** which are well reported and well written publications.
- Included only 3 instances of missing data, limiting the ability to assess the risk for **data fabrication** of the LLM.
- Did not evaluate how the LLM can be **integrated into the workflow of a systemtic review**, e.g. as a complement to human reviewers or as a potential replacement.

Image by Junah Rosales from Pixabay

13

Study within Reviews (SWAR)

- Six **use cases under "real-world"** circumstances of systematic reviews.

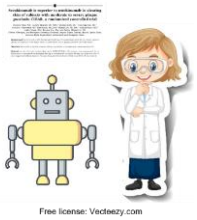
Traditional human-led data extraction

Free license: Vecteezy.com

VS.

- Concordance
- Accuracy
- Time required

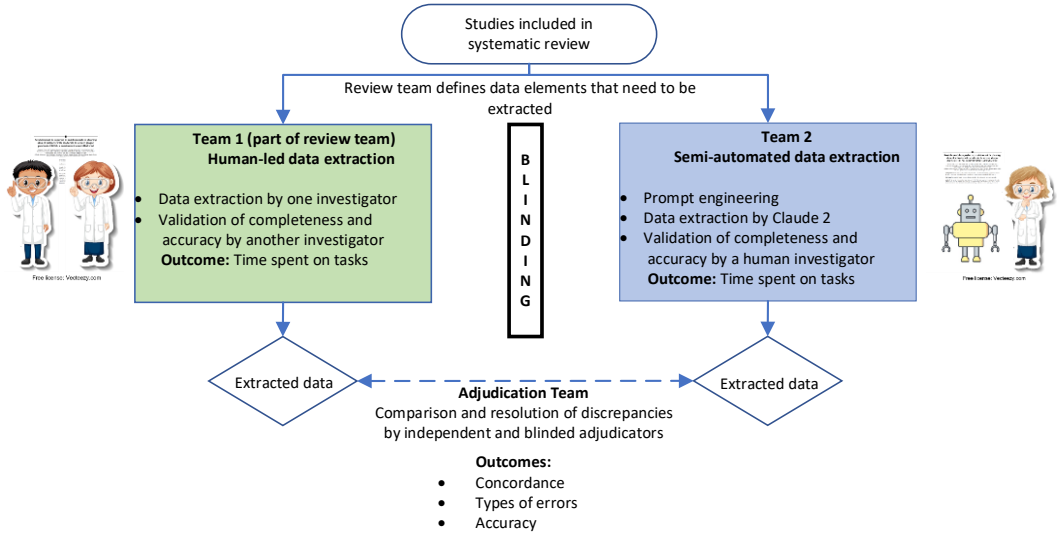
Semi-automated data extraction replacing one human



Free license: Vecteezy.com

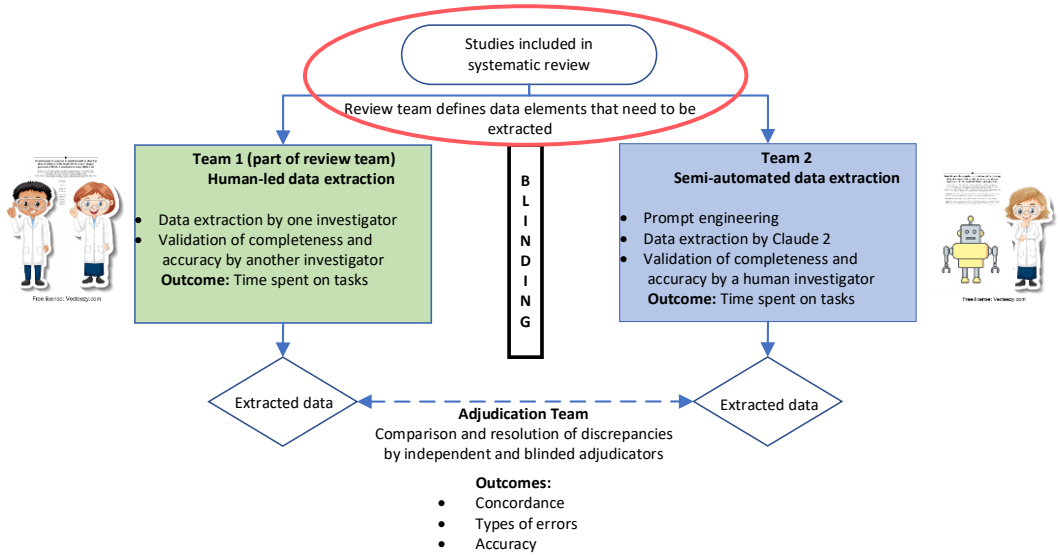
14

Study Design: Prospective Parallel Group Study



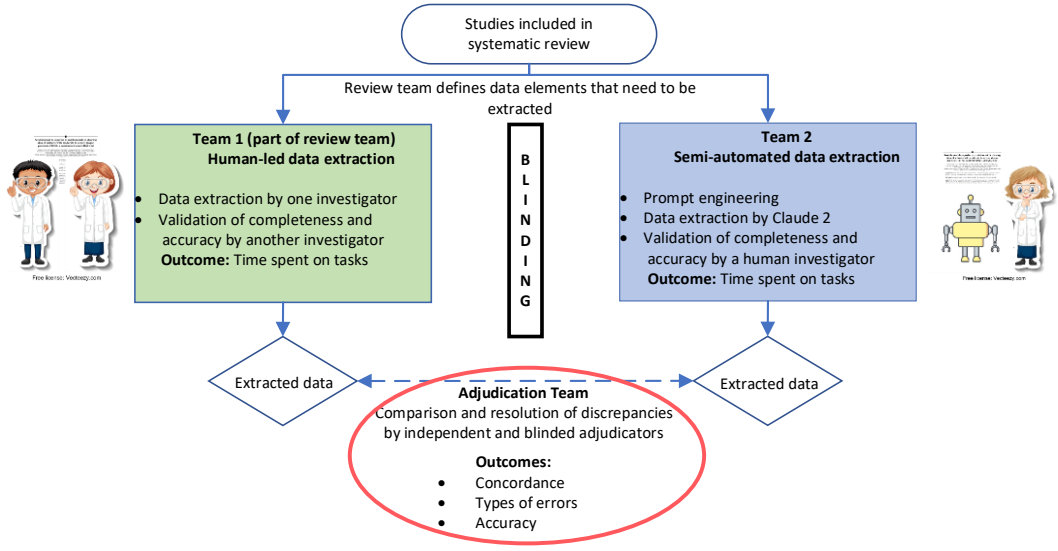
15

Study Design: Prospective Parallel Group Study



16

Study Design: Prospective Parallel Group Study



17

17

Tasks of Adjudicating Team

- Evaluation of performance of the approaches and **classification of errors**.
- For any **discrepancies in extracted data**, adjudicators check the journal publications.
- In cases where data extractions by humans were incorrect, they **revise reference standard**.

Concordance is **factual congruence** of extracted data items, even if there are variations in style, presentation, or length between the two data extractions.

18

18

Who made the mistake?

- Extraction of Team A was incorrect.
- Extraction of Team B was incorrect.
- Both teams were incorrect.
- Neither team was incorrect.
  - Definitions of data elements or prompt language were sometimes vague or ambiguous.
  - E.g., one group extracted ITT results, the other per-protocol results

Severity of Errors

Error	Definitions
Major error	This error significantly compromises the accuracy of the data, and, if uncorrected, could lead to erroneous conclusions.
Minor error	This error is less severe than a major error and may or may not impact interpretation of the existing data.
Inconsequential difference	This difference most likely would not impact the interpretation of the data

Operationalization of Adjudications

First adjudicator:

- Assesses concordance of data extractions and checks original articles
- Assigns error severity ratings.
- Identifies which group made the incorrect extraction.

Second adjudicator:

- Reviews all discordant results and verifies errors severity ratings and group which made mistake

Third adjudicator:

- Resolves discrepancies between the first and second reviewers.

22

Preliminary Findings

23

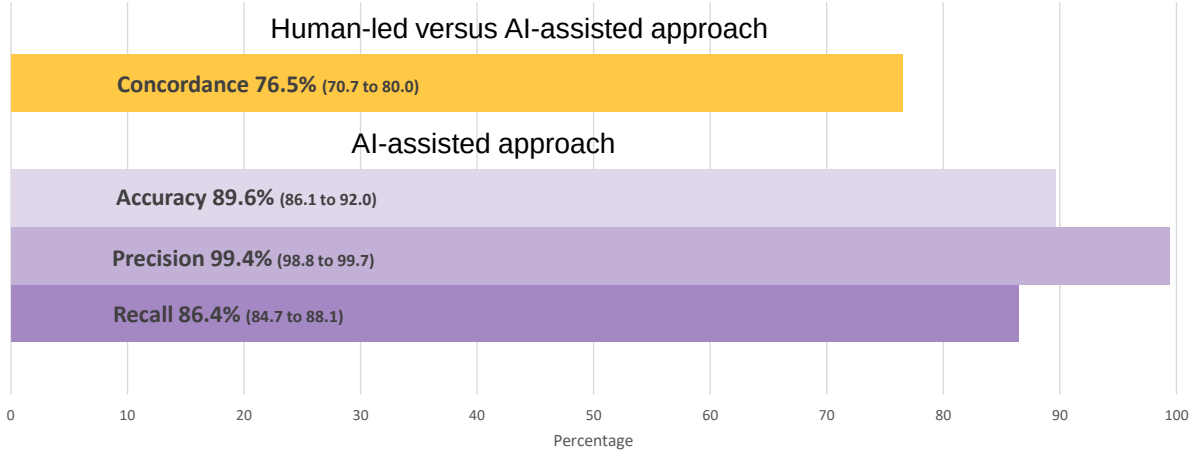
Characteristics of Reviews

Topic	K studies	N data items
Implementation strategies for interventions to prevent mental health disorders in children/adolescents	11*	891
Interventions to Improve Care of Bereaved Persons	20*	1.337
Total	**31**	**2.228**

* Included RCTs and NRSI

24

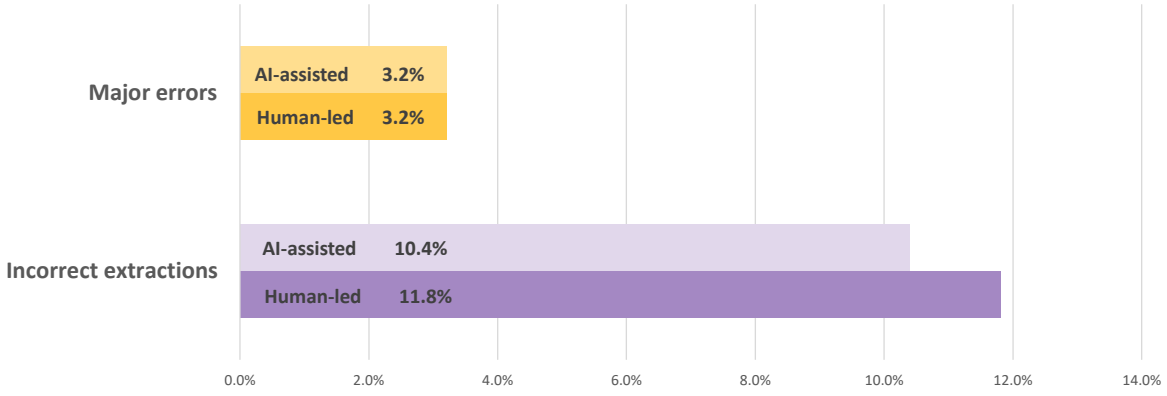
Concordance and Accuracy Metrics



25

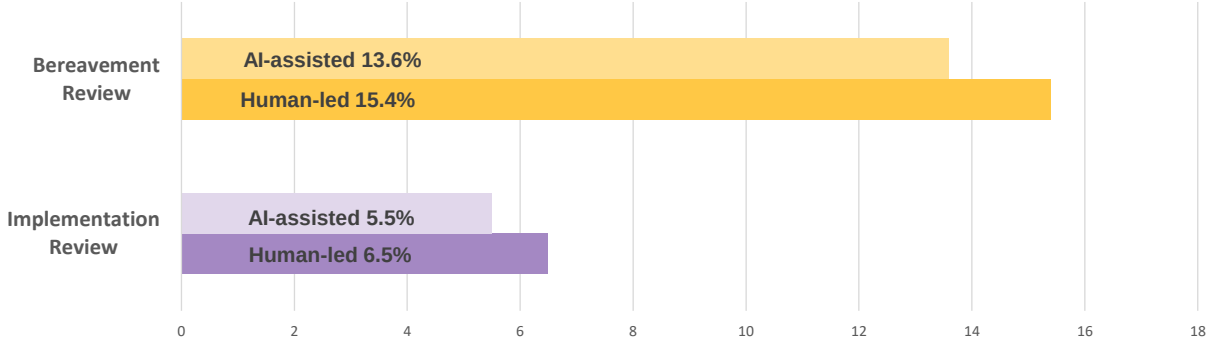
Incorrect Data Extractions and Major Errors

Major error This error significantly compromises the accuracy of the data, and, if uncorrected, could lead to erroneous conclusions.



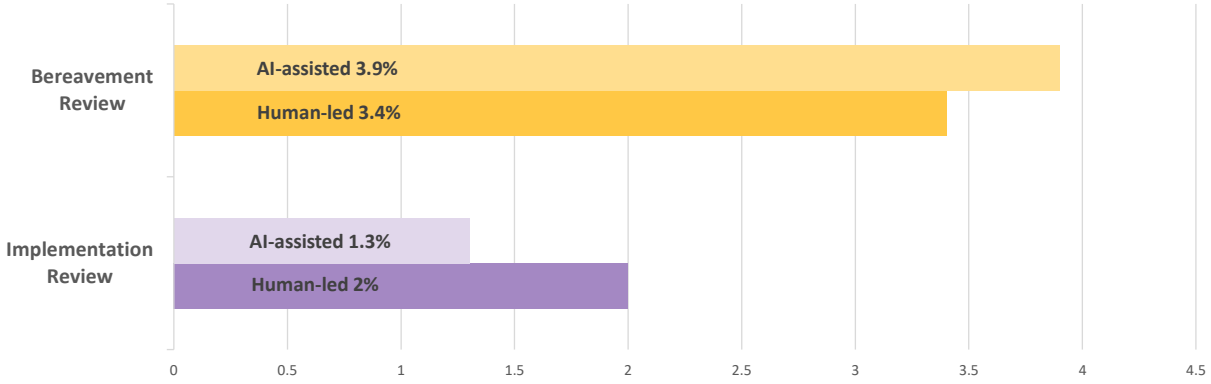
26

Proportions of Incorrect Data Extractions by Review



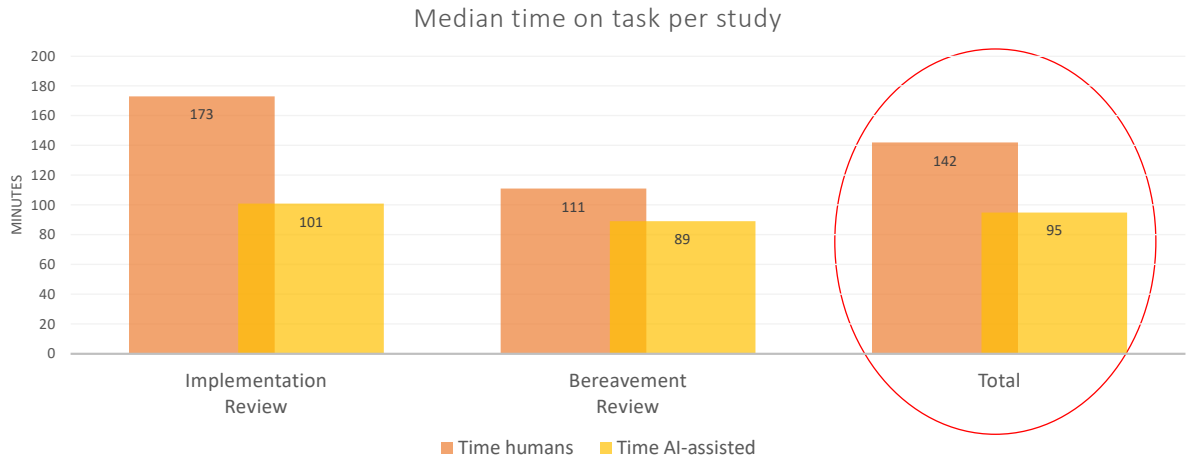
27

Proportions of Major Errors by Review



28

Time Required



29

Limitations and Practical Challenges

- Workflow validation studies have potentially **restricted generalizability**.
- The **choice of the topic** for validation can impact results. Randomized trials may be easier for both humans and machines to accurately extract than non-randomized designs.
- **Human variation** can significantly impact validation studies.
- By the time a study is completed, the LLM under evaluation may have been replaced by a **newer model**.

Methodological Challenges: Humans are an Imperfect Reference Standard

- Human data extraction is as an imperfect reference standard and should not be viewed as a “gold standard”.
- Is some degree of **non-inferiority** for an LLM-assisted data extraction process **acceptable**?

OR

- Should (semi-) automated data extraction not only match but ideally **surpass the performance of human** data extraction?

Challenges: Risk of Data Contamination

- **Sources of data** for training of LLMs often remain unspecified.
- If the model has encountered the data during training, it may "**memorize**" the information, **artificially enhancing performance**.
- The **extent of bias** introduced by data contamination is **not known**.

Image by Tumisu from Pixabay

32

Thank you

ggartlehner@rti.org

33