Can Artificial Intelligence Tools Enhance Data Abstraction During Systematic Literature Reviews?

Kristen A. Cribbs, PhD, MPH, Wesley T. Baisley, Lucas T.A. Blackmore, MPH, Betsy J. Lahue, MPH
Alkemi LLC, Manchester Center, VT, USA

MSR9

alkemi® Unlocking Value, Achieving Access.

Background

- In health economic and outcomes research, systematic literature reviews (SLRs) are integral to obtaining data inputs and assessing the impact of health technologies
- Data abstraction of publication and outcomes data is prone to human error, compromising research integrity. Human error rates in abstraction be as high as 50%, with most estimates around 10-30%[1,2]
- Our study assessed abstraction error using a publicly available large language model artificial intelligence (AI) tool (Microsoft Copilot)

Methods

- 33 publications were identified during an SLR, requiring abstraction of 33 data points across 7 domains for each publication
- 7 AI prompts were developed, tested, and validated for each of the abstraction domains **(Table 1)**
- Study methodology involved executing AI prompts, assessing errors, and conducting analyses **(Figure 1)**
- Descriptive analysis were conducted to calculate error rates overall as well as by abstraction domain, publication, and error type (inaccurate, incomplete)

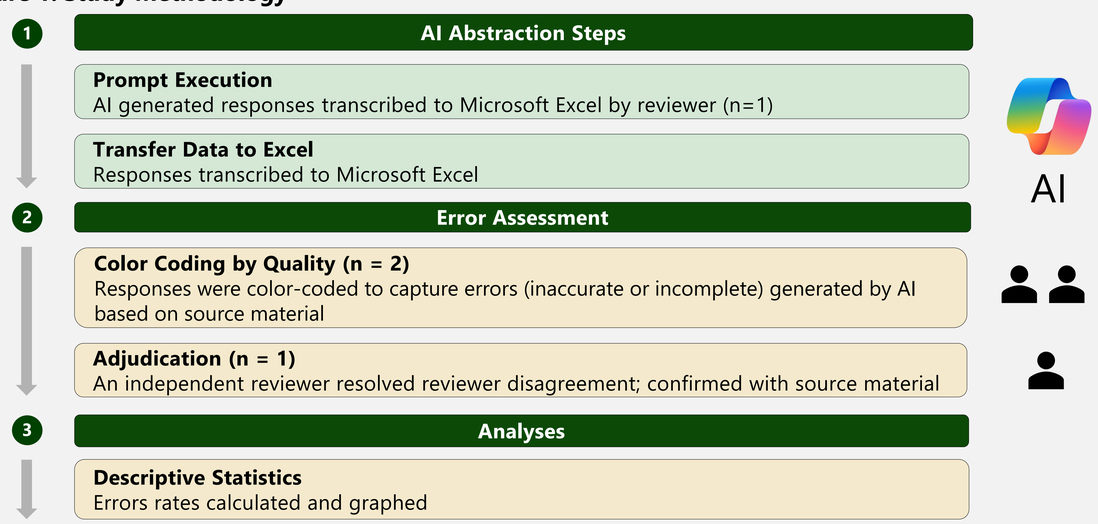
Table 1. AI Prompt Information

Abstraction Domain (n=33 data points)	Parameters
Publication Information (n=7)	Authors, Publication Year, Title, Publication Type, Sponsor or Funding Source, Ethics Approval, Informed Consent
Treatments Studied (n=4)	Treatment Technology Type, Brand, Manufacturer, Comparator(s)
Study Design and Methodology (n=8)	Study Location, Study Design, Sample Size, Inclusion Criteria, Exclusion Criteria, Primary Endpoints, Secondary Endpoints, Follow-Up Period
Baseline Patient Characteristics (n=4)	Prior Therapies, Age, Baseline Staging Score, Baseline Biomarker Level
Treatment Parameters (n=1)	Procedural Parameters
Efficacy Outcomes (n=3)	Percent Patients with Negative Test Result, Percent Decrease in Total Organ Volume, Overall Percent Reduction in Biomarker Level
Safety Outcomes (n=6)	Percent Patients Experiencing an Adverse Event (AE), AE Grading System Used, Percent AEs by Grade Value, Percent Patients Experiencing a Severe AE (SAE), AE Grading System Used, Percent SAE by Grade Value

Sample Prompt:

"Please generate a table with the following information from THIS PAGE (if NOT included put "NR"): Column 1: "Authors" (note all author, format [Last name]. [first name initial]) Column 2: "Publication Year" (year published)…"

Figure 1. Study Methodology



Results

- Execution of the 7 AI prompts yielded a total of 1089 populated data cells for the 33 publications
- The overall AI abstraction error rate was 13% (142/1089) **(Figure 2)**
- Most abstraction errors were inaccuracies (10%, 106/1089) versus incomplete information (3%, 36/1089)
- Error frequency by publication ranged from 0 to 10, with mean 4.3 ± 2.28 errors per publication **(Figure 3)**
- The 'Publication Information' domain had the lowest error rate (2%, 5/231), while 'Efficacy Outcomes' had the greatest (42%, 42/99) **(Figure 4)**
- Stratified findings revealed inaccuracy errors were more prevalent than incompletion errors across all but 2 abstraction domains ('Publication Information,' 'Treatment Parameters') **(Figure 4)**

Figure 2. AI Abstraction Error Rate (n=1089)

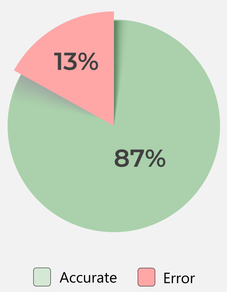
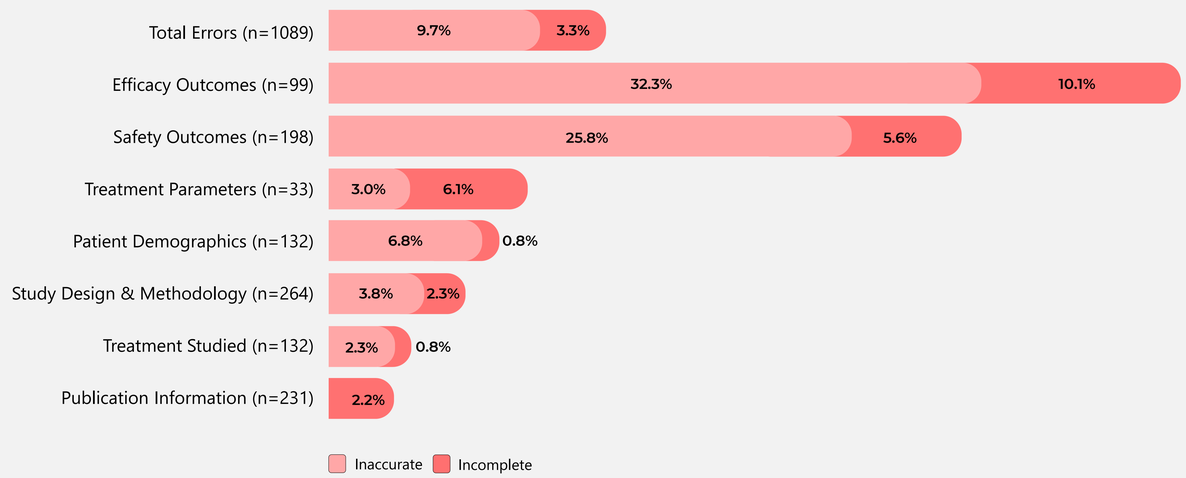


Figure 3. Publication Error Histogram (n=33 articles)

Figure 4. Error Rates Per Abstraction Domain



Conclusions

- To our knowledge, this is the first documented study to use a commercially available large language model for SLR abstraction
- On average, observed AI error rates were lower than published estimates of human error rates
- Of concern, AI abstraction of efficacy and safety outcomes had the highest error rates
- Human quality control is essential to ensure robust and reliable SLR abstraction for all variables

References: 1. E JY, Saldanha IJ, Canner J, Schmid CH, Le JT, Li T. Adjudication rather than experience of data abstraction matters more in reducing errors in abstracting data in systematic reviews. Res Synth Methods. 2020;11(3):354-362. doi:10.1002/jrsm.1396 **2.** 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. Published 2017 Nov 28. doi:10.1186/s12874-017-0431-4

Abbreviations: SLR, systematic literature review; AI, artificial intelligence; AE, adverse event; SAE, Severe adverse event; NR, not reported

Presented at ISPOR, May 2024, Atlanta, Georgia. Please contact betsy.lahue@alkemihealth.com for more information.
www.alkemihealth.com