July 2025

White Paper

Bridge
Evidence that matters

Artificial Intelligence in Systematic Literature Reviews
Part 5 | AI-aided full data extraction

In three earlier papers in this series on AI in systematic literature reviews (SLRs), we reported strong AI performance in title/abstract screening, full-text screening, and methodology extraction. In this fourth paper, we evaluate model performance for full data extraction – one of the most resource-intensive SLR stages. Across 10 SLRs and ~50,000 data points, the model achieved very high accuracy (98-100%) and strong completeness (85-98%). AI performance was consistent across clinical and real-world studies, though some complex data (e.g. subgroup efficacy or risk factors) were occasionally missed. These results support the use of AI for first-draft extraction, combined with full human quality control.

Introduction

Following our previous research demonstrating robust performance of AI in systematic literature review (SLR) stages, including title and abstract (TiAB) screening, full-text screening (FTS), and methodology extraction (or PICOS+ extraction)[1,2,3], this research paper describes our evaluation of AI model performance in **full data extraction.**

Data extraction is crucial in an SLR because it ensures that key information needed for analysis and comparison across studies is consistently and accurately captured. This process reduces bias, supports reproducibility, and forms the basis for meaningful qualitative or quantitative synthesis.

For a typical SLR extracted into an Excel data extraction grid, the total number of columns (i.e., variables to be extracted) can range from ~100 to 200 columns depending on the depth of extraction needed. Full extraction is one of the most resource intensive steps of an SLR and typically occupies around 30% of the total effort.

Our objective was to determine the performance of AI in full data extraction from full-text publications.

Methodology

The scope of our testing included **10 SLRs** previously completed by experienced researchers at Bridge to 'gold-standard' quality, i.e., with 2 reviewers, a third adjudicator, final sign off by a senior colleague at Bridge and acceptance by the client.

The selected ten SLRs were all of different indications and comprised 5 clinical trial-focused SLRs and 5 real-world evidence (RW) SLRs.

From each SLR, we randomly selected 10 publications (including any associated supplementary materials), resulting in a validation dataset of 100 publications. Collectively, this dataset covered approximately **50,000 individual data points** for machine extraction.

For each variable, we developed bespoke prompts, which were first tested on publications not included in the validation dataset. We then used these prompts to extract data from the 100 publications using the OpenAI o3-mini model.

Importantly, the AI team conducted their extractions without any visibility of the 'gold-standard' extraction database. Those databases were accessed only **after** the AI team had completed their work, for the purposes of comparing AI outputs v human 'gold standard' outputs.

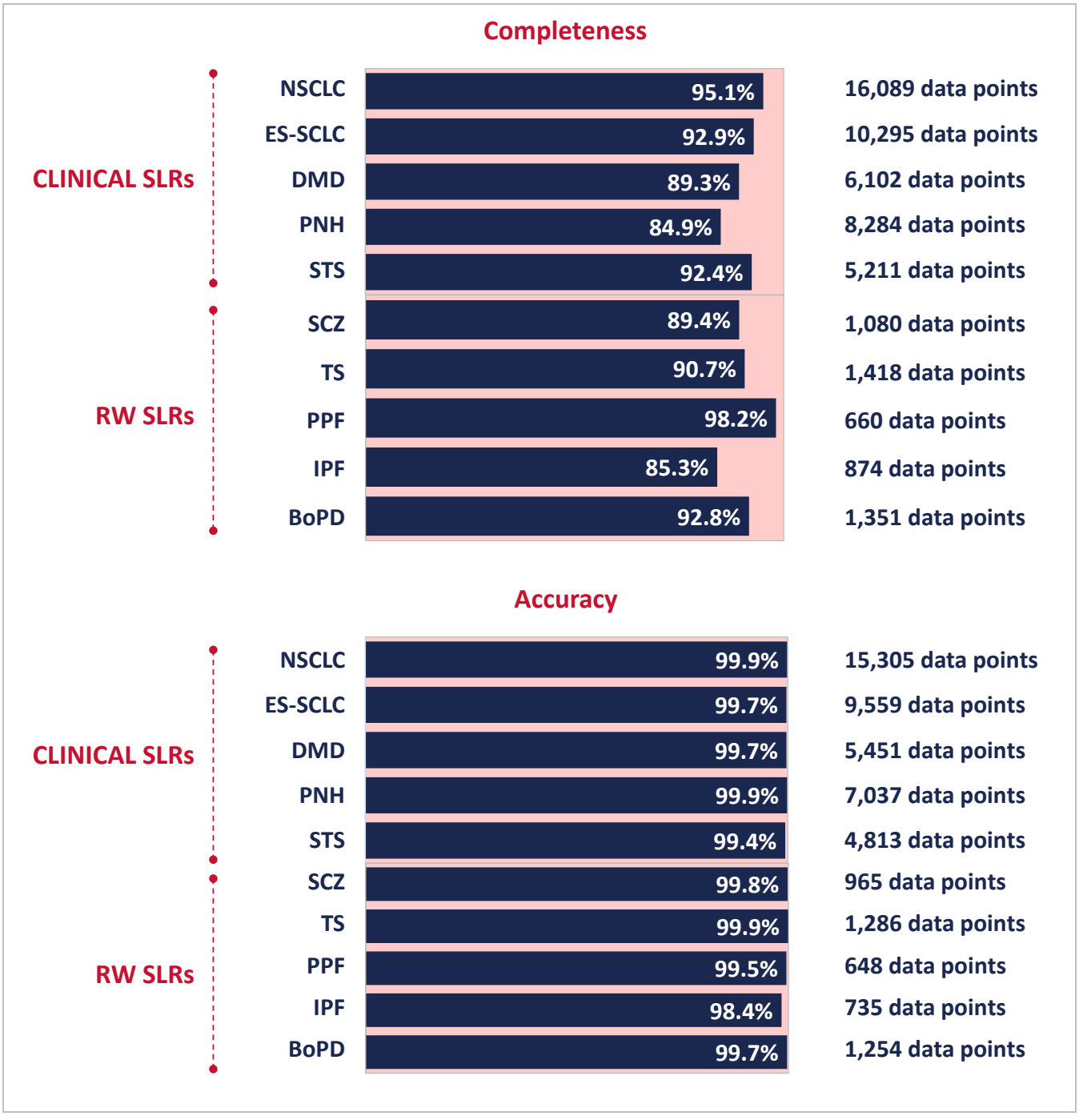
We evaluated AI performance using the following metrics:

1. **Completeness:** Defined as the extent to which AI extracted data for the relevant variables, across all subgroups of interest, such as age, gender, line of treatment, and others. 100% completeness would imply that no data was missing.
2. **Accuracy:** For the AI-extracted data, defined as correct match between each data element extracted by the AI and the corresponding extractions in the 'gold-standard' dataset.

Results

We were able to demonstrate high completeness levels and consistently very high accuracy for all variables across the 10 SLRs using o3-mini (**Figure 1**). The completeness ranged from **85% to 98% (median 92%),** and the accuracy ranged from **98% to 100% (median 100%),** across 10 SLRs covering a total of **51,352 data points**.

Figure 1: Summary results for completeness* and accuracy of data extraction across the 10 SLRs



*Completeness: The extent to which AI extracted data for all the relevant variables.
Accuracy: The proportion of AI-extracted data that was extracted correctly (i.e., matched with the 'gold-standard' dataset).

Note that the number of data points used to calculate accuracy is lower than for completeness, as accuracy was assessed only on data points fully extracted by AI.

BoPD = Borderline Personality Disorder; **DMD** = Duchenne Muscular Dystrophy; **ES-SCLC** = Extensive-Stage Small Cell Lung Cancer; **IPF** = Idiopathic Pulmonary Fibrosis; **NSCLC** = Non-Small Cell Lung Cancer; **PNH** = Paroxysmal Nocturnal Haemoglobinuria; **PPF** = Progressive Pulmonary Fibrosis; **RW** = Real world; **SCZ** = Schizophrenia; **SLR** = Systematic Literature Review; **STS** = Soft Tissue Sarcoma; **TS** = Turner Syndrome.

The actual completeness and accuracy results for each variable assessed in the 10 SLRs are provided in **Tables 1** and **2.** The key findings are summarised after the tables.

Table 1: Completeness of data extraction across the 10 SLRs

	Variable ↓ SLR →	STS	PNH	DMD	ES-SCLC	NSCLC
Clinical trial SLRs	Study characteristics	100%	98%	91%	100%	100%
	Patient baseline characteristics	100%	91%	82%	100%	100%
	Treatment details	94%	100%	100%	100%	100%
	Efficacy	88%	72%	88%	89%	92%
	Safety – Overall	100%	100%	100%	100%	100%
	Safety – Specific adverse events		93%	94%		
	Discontinuation	85%	100%	100%	100%	
	Variable ↓ SLR →	**BoPD**	**IPF**	**PPF**	**TS**	**SCZ**
Real-world study SLRs	Methods	99%	96%	96%	97%	97%
	Incidence		100%	91%		
	Incidence/ prevalence				94%	
	Prevalence	60%	100%	100%		
	Diagnostic techniques		100%	100%		
	Clinical features	68%	97%	100%	89%	52%
	Clinical features [Factors]	100%			98%	
	Clinical subgroups		100%			
	Age at diagnosis		85%	100%	98%	
	Age at presentation				100%	
	Comorbidities	100%	100%	100%		100%
	Comorbidity [Factors]	100%				
	Acute exacerbation		100%	100%		
	Acute exacerbation [Factors]		56%			
	Natural history/ progression	100%	42%	71%	91%	89%
	Natural history [Factors]	100%	32%		23%	
	Transplant-free survival		100%	100%		
	Mortality	100%	74%		99%	
	Mortality [Factors]		68%		100%	
	HRQoL	100%	100%		86%	100%
	HRQoL [Factors]	100%			29%	
	Activities of daily living	100%	44%	100%		89%
	Activities of daily living [Factors]	0%				65%
	Occupational functioning	77%				
	Occupational functioning [Factors]	85%				
	Caregiver burden	91%				
	Caregiver burden [Factors]	100%				
	Treatment adherence	100%	100%	100%		
	Treatment patterns	100%	93%	100%		100%
	Treatment patterns [Factors]	100%				
	Dose reduction		100%	100%		
	Healthcare resource utilisation	88%			100%	
	Healthcare resource utilisation [Factors]	100%			100%	
	Hospitalisation		77%			
	Direct costs	100%			100%	

BoPD = Borderline Personality Disorder; **DMD** = Duchenne Muscular Dystrophy; **ES-SCLC** = Extensive-Stage Small Cell Lung Cancer; **HRQoL** = Health-Related Quality of Life; **IPF** = Idiopathic Pulmonary Fibrosis; **NSCLC** = Non-Small Cell Lung Cancer; **PNH** = Paroxysmal Nocturnal Haemoglobinuria; **PPF** = Progressive Pulmonary Fibrosis; **SCZ** = Schizophrenia; **SLR** = Systematic Literature Review; **STS** = Soft Tissue Sarcoma; **TS** = Turner Syndrome.
Light grey cells represent variables that were not relevant for the selected studies in the testing dataset for a particular SLR, and which therefore had not been extracted for that review.

Legend:

100%	≥90% to 99%	<90%

Table 2: Accuracy of data extraction across the 10 SLRs

	Variable ↓ SLR →	STS	PNH	DMD	ES-SCLC	NSCLC
Clinical trial SLRs	Study characteristics	98%	97%	97%	96%	98%
	Patient baseline characteristics	99%	100%	100%	100%	100%
	Treatment details	99%	99%	98%	98%	98%
	Efficacy	100%	100%	100%	100%	100%
	Safety – Overall	99%	99%	100%	100%	100%
	Safety – Specific adverse events		100%	100%		
	Discontinuation	100%	100%	100%	100%	
	Variable ↓ SLR →	**BoPD**	**IPF**	**PPF**	**TS**	**SCZ**
Real-world study SLRs	Methods	99%	99%	98%	99%	99%
	Incidence		100%	100%		
	Incidence/ prevalence				100%	
	Prevalence	100%	100%	100%		
	Diagnostic techniques		100%	100%		
	Clinical features	100%	100%	100%	100%	100%
	Clinical features [Factors]	100%			100%	
	Clinical subgroups		100%			
	Age at diagnosis		59%	100%	100%	
	Age at presentation				100%	
	Comorbidities	97%	100%	100%		100%
	Comorbidity [Factors]	100%				
	Acute exacerbation		100%	100%		
	Acute exacerbation [Factors]		100%			
	Natural history/ progression	100%	100%	100%	100%	100%
	Natural history [Factors]	100%	100%		100%	
	Transplant-free survival		84%	100%		
	Mortality	100%	100%		100%	
	Mortality [Factors]		100%		100%	
	HRQoL	100%	100%		100%	100%
	HRQoL [Factors]	100%			100%	
	Activities of daily living	100%	100%	100%		100%
	Activities of daily living [Factors]	100%				100%
	Occupational functioning	100%				
	Occupational functioning [Factors]	100%				
	Caregiver burden	100%				
	Caregiver burden [Factors]	100%				
	Treatment adherence	100%	100%	100%		
	Treatment patterns	100%	100%	100%		100%
	Treatment patterns [Factors]	100%				
	Dose reduction		100%	100%		
	Healthcare resource utilisation	100%			100%	
	Healthcare resource utilisation [Factors]	100%			100%	
	Hospitalisation		100%			
	Direct costs	100%			100%	

BoPD = Borderline Personality Disorder; **DMD** = Duchenne Muscular Dystrophy; **ES-SCLC** = Extensive-Stage Small Cell Lung Cancer; **HRQoL** = Health-Related Quality of Life; **IPF** = Idiopathic Pulmonary Fibrosis; **NSCLC** = Non-Small Cell Lung Cancer; **PNH** = Paroxysmal Nocturnal Haemoglobinuria; **PPF** = Progressive Pulmonary Fibrosis; **SCZ** = Schizophrenia; **SLR** = Systematic Literature Review; **STS** = Soft Tissue Sarcoma; **TS** = Turner Syndrome.
Light grey cells represent variables that were not relevant for the selected studies in the testing dataset for a particular SLR, and which therefore had not been extracted for that review.

Legend:

100%	≥90% to 99%	<90%

Completeness

Overall, the o3-mini model delivered reasonably strong performance in terms of completeness of data extracted across the 10 SLRs (**Table 1**). For clinical SLRs, the overall completeness ranged from **85% to 95% (median 92%),** and for RW SLRs, it ranged from **85% to 98% (median 91%)** across reviews**.** In terms of individual variables, for 62 variables, 100% of the data were fully extracted; for 21 variables, the completeness was 90-100%; and for 29 variables, the completeness was <90%. Some examples of challenges with data completeness included:

- For efficacy parameters, completeness was sometimes low when a paper reported multiple outcomes across several subgroups.
- In review articles presenting pooled estimates, the model frequently returned only the pooled estimates, despite being instructed to extract study-specific data as well.
- The model sometimes failed to extract non-significant results for outcome-associated factors despite being explicitly prompted to do the same.

While the model occasionally missed relevant data in more complex cases – particularly for data-heavy extractions such as efficacy outcomes and factors associated with outcomes – it was reassuring that it achieved complete (i.e., 100%) data extraction for over half of the variables.

Accuracy

When the model did extract data, the accuracy of that extracted data was consistently very high across both clinical and RW SLRs, and across all variable types (**Table 2**). In clinical SLRs, overall accuracy ranged from **99% to 100% (median 100%)**, while in RW SLRs, it ranged from **98% to 100% (median 100%)** across reviews. In terms of individual variables, accuracy was 100% for 90 variables, between 90–100% for 19 variables, and below 90% for only two variables.

Reflections and next steps

In our testing, AI-assisted full data extraction in SLRs demonstrated strong performance. Across approximately 50,000 data points from 10 SLRs – covering both clinical trial and RW studies – the o3-mini model achieved a median completeness rate of 92% and a median accuracy rate approaching 100%. Performance was consistent across both clinical and RW SLRs.

These findings have two key implications:

1. **AI is now reliable enough to be used routinely for first-draft data extraction**

   With strong accuracy rates on extraction across multiple SLRs, there is supporting evidence that appropriate use of AI can improve efficiency and quality of extraction when in a combined workflow with expert humans.

2. **Despite the high accuracy, completeness issues and rare errors persist, making 100% human QC essential.**

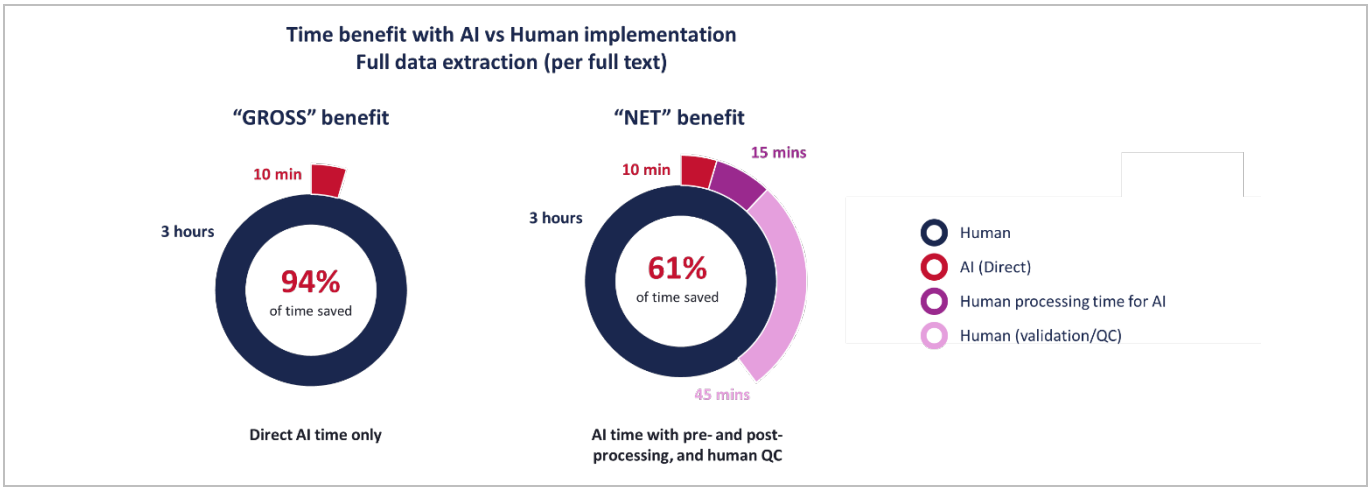
   Based on our findings, AI-extracted data is currently inadequate without expert human QC. When the model entirely skips relevant data, any downstream analysis would be incomplete and potentially misleading.

   Moreover, while the overall inaccuracy rate is low as a proportion of the total data points assessed, the errors are not always trivial. Across the 10 SLRs covering approximately 50,000 data points, 130 inaccuracies were identified – about 0.28% of the total. Inaccuracies are more concerning because they involve data cells where the model has **confidently entered incorrect values** – this poses a greater risk to data integration and further analysis than missing data.

   Therefore, for now, rigorous human QC remains critical. This aligns with the "human-in-the-loop" approach endorsed by NICE and other researchers in the field, ensuring both the reliability and integrity of the final dataset.[4-7]

The use of AI for first-pass data extraction – **combined with 100% human quality control** – **has important implications for the future of SLRs. Figure 2** illustrates the efficiency gains achieved through this approach, based on internal resource metrics. When considering only the direct time taken by the AI model to perform full data extraction, gross time savings reached **94%.** However, in practice, additional steps – such as pre-processing inputs, post-processing outputs, and conducting human QC – need to be factored in. After accounting for these, **net time savings remain substantial at 61%,** highlighting a substantial efficiency gain for what is typically one of the most resource-intensive stages of an SLR.

Figure 2: Time benefit with AI vs Human implementation for full data extraction



AI= artificial intelligence; QC=quality control

Looking ahead, we might reasonably expect AI performance to continue improving**.** While accuracy is already near – but not yet at – ceiling levels, we can reasonably expect completeness metrics to improve with ongoing model advancements and better prompt design.

Our findings here consistently show that the **AI + single human QC** approach is both **faster and more resource-efficient** than traditional manual methods.

We will shortly be publishing additional white papers on our research findings on [a] the role of AI in table narratives & [b] the role of AI in critical appraisal in the context of SLRs.

Authors: Saifuddin Kharawala, Pankdeep Chhabra, Divyanshu Jindal, and Paul Gandhi

References

1. Kharawala S, Issacs S, Jindal D, Gandhi P. Artificial Intelligence in Systematic Literature Reviews Part 1 | AI-aided Title/Abstract Screening. Available at: https://www.bridgemedical.org/site/assets/files/2156/ai_in_literature_reviews_white_paper_31_jan_24-1.pdf.

2. Kharawala S, Issacs S, Chhabra P, Jindal D, Gandhi P. Artificial Intelligence in Systematic Literature Reviews Part 2 | AI-aided Full-text screening. Available at: https://www.bridgemedical.org/site/assets/files/2202/white_paper_on_ai-aided_full-text_screening_in_slrs_23_august_2024-1.pdf.

3. Kharawala S, Chhabra P, Jindal D, Gandhi P. Artificial Intelligence in Systematic Literature Reviews Part 4 | AI-enabled initial data extraction. Available at: https://www.bridgemedical.org/site/assets/files/2230/white_paper_on_ai-enabled_initial_data_extraction.pdf. Accessed 6th April 2025.

4. Use of AI in evidence generation: NICE position statement. Available at: https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement. Accessed 11th July 2025.

5. Schmidt L, Hair K, Graziozi S, et al. Exploring the use of a Large Language Model for data extraction in systematic reviews: a rapid feasibility study. Proceedings of the 3rd Workshop on Augmented Intelligence for Technology-Assisted Reviews Systems, 2024, arXiv:2405.14445.

6. Sun Z, Zhang R, Doi SA, et al. How good are large language models for automated data extraction from randomized trials? medRxiv 2024.02.20.24303083.

7. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. *Res Synth Methods*. Published online June 19, 2024. doi:10.1002/jrsm.1732.