March 2025

White Paper

Bridge
Evidence that matters

Artificial Intelligence in Systematic Literature Reviews
Part 4 | AI-enabled Initial Data Extraction

In the first two papers in this series on AI in systematic literature reviews (SLRs), we presented our methodology and results for testing the performance of AI models in title/abstract (TiAB) and full-text screening (FTS), with which we achieved high sensitivity (up to 96% for TiAB screening, and ≥99% for FTS). In this paper, we report our methodology and findings on the subsequent SLR step, **initial data extraction**, using the OpenAI o1-mini model. Overall, the model demonstrated consistently high accuracy and sensitivity (≥90% for most variables) across our validation datasets. These results also indicate the model is likely to perform very well in full-scale extraction (the subject of our next white paper due to be released within the next few weeks).

Introduction

Extending our earlier work on the use of AI in systematic literature reviews (SLRs)[1,2], this research paper describes our findings on the use of AI models for initial data extraction (known variously by clients as 'categorisation', 'data landscape' or 'population, intervention, comparator, outcomes and study design [PICOS]+ extraction'). To characterise the selected studies after title/abstract (TiAB) screening and full-text screening (FTS), key information is extracted on study methods (e.g., study design, sample size, follow-up duration) and the availability of relevant outcomes data (categorised Yes or No). Although this is not a 'formal' step in an SLR, it is almost always requested by clients to give them an early overview (or 'landscape') of data availability.

It is primarily an extraction task – an extraction of methods rather than results – and operationally is a sub-set of full extraction. During the initial data extraction stage, approximately 15 to 30 data fields (encompassing methodological and outcome variables) are extracted from each full-text publication. This stage may therefore require thousands of extraction fields per SLR.

At the time this validation work was carried out, the latest model from Open AI was the o1-mini model (o1-mini), and so our primary objective was to determine the performance of this model for initial data extraction from full-text publications. We had previously tested initial data extraction with an earlier AI model (GPT-4o), so we also compared the results with o1-mini against those with GPT-4o.

Methodology

We followed a similar protocol as for earlier stages of our research program (AI-enabled TiAB screening and FTS).[1,2] During testing, we developed bespoke prompts for each variable, and applied them across different types of reviews, including three clinical trial–focused and two real-world (RW) study–focused SLRs, each with a different indication to ensure comprehensive coverage. The AI model's performance was compared against our in-house reference datasets for these SLRs – these are 'gold-standard' datasets, since the completed SLRs had been double-human-extracted, QC'd, adjudicated as needed, and reviewed and approved by senior Bridge team members as well as by the end client in each case.

Across the five SLRs used for validation, there were a total of 311 full-text publications (208 from clinical trial-focused SLRs and 93 from RW study-focused SLRs). The initial data extraction encompassed both methodology and outcome variables. For the methodology variables, categories were organised using the PICOS framework (excluding outcomes), along with an additional 'other' category. For the outcome variables, categories included efficacy/safety/discontinuation, treatment, clinical burden, humanistic burden, and economic burden. At this stage of extraction, **the focus was to establish whether an outcome was included** [not the full results], and only 'yes/no' responses were required.

Table 1 shows the range of parameters extracted across the five SLRs, but please note – not all parameters were relevant for each SLR [see **Table 2** for details per SLR]. A total of 5,695 extraction fields in total were included in the initial extraction.

In addition to the specific data extracted, the AI models were also prompted to *provide additional context or a rationale* for each extraction; this rationale helped us refine and improve our prompts while developing bespoke prompts, and the additional context in the rationales was used for quality control checking of the extracted data.

Table 1: Overview of extracted fields for each SLR* during 'Initial Extraction'

Methodology variables	
Variable category	**Variables**
Population	Adult/paediatric population, adjuvant/ neoadjuvant therapy, current line of therapy, previous line of therapy, disease name, dialysis status and disease stage
Intervention	Intervention name and intervention class
Comparator	Comparator
Study design	Review/primary study, study design and study phase
Other	Overall sample size, subgroup (disease-specific) sample size, follow-up period, country, study/trial name, NCT ID, journal article/conference abstract and language

Outcome variables	
Variable category	**Variables**
Efficacy/ Safety/ Discontinuation	Efficacy, safety, discontinuation and tolerability
Humanistic burden	Health-related quality of life, patent-reported symptoms, activities of daily living, work disability and caregiver burden
Clinical burden	Incidence, prevalence, method of diagnosis as per guidelines, demographic characteristics, clinical characteristics, sociodemographic and clinical characteristics, natural history, risk factors/predictors of disease, predictors/risk factors of long-term outcome, comorbidities and mortality
Treatment	Treatment guidelines, treatment effectiveness, treatment patterns, treatment barriers, treatment adherence/compliance and treatment satisfaction
Economic burden	Cost, direct cost, indirect cost, healthcare resource use and health-state utilities

* The number of extraction fields in the SLRs were: 1,683 for soft tissue sarcoma, 1,391 for anaemia-CKD, 168 for metastatic castration resistant prostate cancer, 928 for borderline personality disorder, and 1,525 for atopic dermatitis.

We evaluated AI performance as per the following metrics:

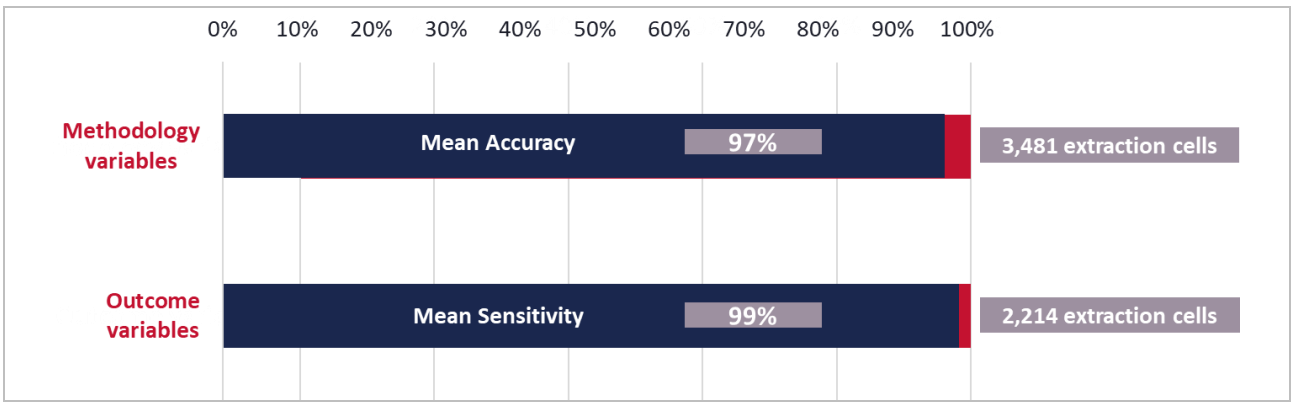
1. **Accuracy:** This applied to methodology variables and was defined as the proportion of correct matches between the data extracted by AI and that extracted by the humans in the reference datasets.
2. **Sensitivity:** This applied to outcome variables categorised as 'Yes' or 'No' and was defined as the proportion of actual "Yes" instances that were correctly classified by AI as "Yes". This metric emphasises measurement of false negatives, i.e., instances of AI overlooking available data. Note that at the initial extraction stage, we are only identifying the presence of an outcome measure, not the actual results data (this is the subject of full extraction).

For each variable, the accuracy/sensitivity was classified as *high* if it was ≥90%, moderate if ≥80% to <90%, and *low* if <80%.

Results

We were able to demonstrate consistently high accuracy for the methodology variables and high sensitivity for outcome variables across the five SLRs using o1-mini. For the **methodology variables**, **the mean accuracy was 97% and the median accuracy was 99% (range 85%-100%)** across the different variables tested (**Figure 1**). Similarly, for **outcome variables**, **the mean sensitivity was 99% and the median sensitivity was 100% (range 88%-100%)** (**Figure 1**).

Figure 1: Summary results for accuracy* for methodology variables and sensitivity# for outcome variables



*Accuracy: The proportion of correct matches between the data extracted by AI and by the humans in the reference datasets.
#Sensitivity: The proportion of actual "Yes" instances in the reference datasets that were correctly classified by AI as "Yes".

The actual accuracy and sensitivity results for each variable assessed in the five SLRs are provided in **Table 2** and **Table 3** below. The key findings are summarised after each table.

Table 2: Accuracy* for **methodology variables** across the five SLR projects assessed using o1-mini model

Variable category	Variables	Clinical trial-focused SLRs			RW study-focused SLRs	
		STS (n=99)	Anemia-CKD (n=107)	mCRPC (n=12)	BOPD (n=32)	AD (n=61)
Population	Adult/Paediatric population	99%	100%			
	Adjuvant/Neoadjuvant therapy	92%				
	Current line of therapy	97%				
	Previous line of therapy			90%		
	Disease name	93%		100%		
	Dialysis status		98%			
	Disease stage	100%	85%			
Intervention	Intervention name	99%	94%	100%		
	Intervention class		98%			
Comparator	Comparator name	97%		100%		
Study design	Review/Primary study				100%	100%
	Study design	100%	99%	100%	100%	90%
	Study phase			92%		
Other	Overall sample size	90%	98%	100%	97%	97%
	Subgroup (disease-specific) sample size		96%		96%	100%
	Follow-up period		91%		97%	98%
	Country	100%	96%		100%	100%
	Study/Trial name	92%	100%	100%		
	NCT ID	100%	99%	100%		
	Journal article/Conference abstract	100%	100%	100%	100%	100%
	Language of the publication				100%	
Mean accuracy across variables		**97%**			**98%**	
Median accuracy across variables		**99%**			**100%**	
Range (Min-Max) of accuracy across variables		**85% to 100%**			**90% to 100%**	

*Accuracy: The proportion of correct matches between the data extracted by AI and by the humans in the reference datasets.
n=number of publications tested in each SLR project
AD=atopic dermatitis; BoPD=borderline personality disorder; CKD=chronic kidney disease; mCRPC=metastatic castration-resistant prostate cancer; SLR=systematic literature review; STS=soft tissue sarcoma.
Light grey cells represent variables that were not relevant for a particular SLR, and which therefore had not been extracted for that review. E.g., intervention and comparator were not in the two RW study-focused SLRs.

Legend: Cut-offs for accuracy:

100%	≥90% to 99%	<90%

The o1-mini model delivered robust performance in extracting **methodology variables** across five SLRs, **surpassing 90%** accuracy for most variables. For instance, for all five SLRs, the accuracy was 100% for extracting data on journal article/conference abstract, while it ranged from 90% to 100% for study design and overall sample size. Only one variable fell below 90%, namely, disease 'staging' in the anaemia-chronic kidney disease (CKD) literature review; with the AI model extracting this information incorrectly in approximately 15% of the cases. Interestingly, data on disease staging were 100% accurate in the soft tissue sarcoma (STS) review.

Table 3: Sensitivity* for **outcome variables** across the five SLR projects assessed using o1-mini model

Variable category	Variables	Clinical trial-focused SLRs		RW study-focused SLRs	
		STS (n=99)	mCRPC (n=12)	BOPD (n=32)	AD (n=61)
Efficacy/Safety/ Discontinuation	Efficacy	100%	100%		
	Safety	100%	100%		
	Discontinuation	98%			
	Tolerability		100%		
Humanistic burden	Health-related quality of life	100%		100%	100%
	Patent-reported symptoms				94%
	Activities of daily living			100%	100%
	Work disability			100%	
	Caregiver burden			100%	100%
Clinical burden	Incidence				100%
	Prevalence			100%	100%
	Method of diagnosis as per guidelines			100%	
	Demographic characteristics			100%	
	Clinical characteristics			100%	
	Sociodemographic and clinical characteristics				100%
	Natural history			100%	88%
	Risk factors/predictors of disease				
	Predictors/risk factors of long-term outcome			100%	100%
	Comorbidities			100%	100%
	Mortality			89%	100%
Treatment	Treatment guidelines			100%	100%
	Treatment effectiveness			100%	
	Treatment patterns			89%	
	Treatment barriers			100%	
	Treatment adherence/compliance				100%
	Treatment satisfaction				100%
Economic burden	Cost			100%	
	Direct cost				100%
	Indirect cost				90%
	Healthcare resource use			100%	90%
	Health-state utilities			100%	100%
Mean sensitivity across variables		**100%**		**98%**	
Median sensitivity across variables		**100%**		**100%**	
Range (Min-Max) of sensitivity across variables		**98% to 100%**		**88% to 100%**	

*Sensitivity: The proportion of actual "Yes" instances in the reference datasets that were correctly classified by AI as "Yes". Sensitivity could not be calculated for three variables (Incidence and treatment adherence/compliance for BoPD and risk factors/predictors of disease for AD) due to zero true positives.
n=number of publications tested in each SLR project
AD=atopic dermatitis; BoPD=borderline personality disorder; mCRPC=metastatic castration-resistant prostate cancer; SLR=systematic literature review; STS=soft tissue sarcoma.
Light grey cells represent variables that were not relevant for a particular SLR, and which therefore had not been extracted for that review. E.g., cost and caregiver burden were not assessed in the two clinical trial-focused SLRs.
For Anaemia-CKD, there were no relevant outcome variables for this analysis, and hence that SLR is not included in this table.

Legend: Cut-offs for sensitivity:

100%	≥90% to 99%	<90%

The o1-mini model consistently exhibited high sensitivity in initial data extraction of **outcome variables** across the five SLRs. Outcomes spanned a broad range of domains, from efficacy and safety to humanistic, clinical, and economic burden parameters, which underscores the robustness of the model. The sensitivity was **100% for most variables** and <100% for 7 variables (natural history of disease, mortality, treatment patterns, patient-reported outcomes, indirect costs, healthcare resource use, and discontinuation). For example, the AI model could not identify the data reported on indirect cost (i.e., lost work-days) in an RW study conducted in patients with atopic dermatitis (AD). For another RW study, the AI model could not identify the data reported on treatment patterns (i.e., psychotherapy, group therapy, self-help groups, and residential treatment) in patients with borderline personality disorder (BoPD). It is important to note that across 2,214 extraction points taken together across five SLRs, there were only 7 false negatives (1 each for each of the 7 variables mentioned above).

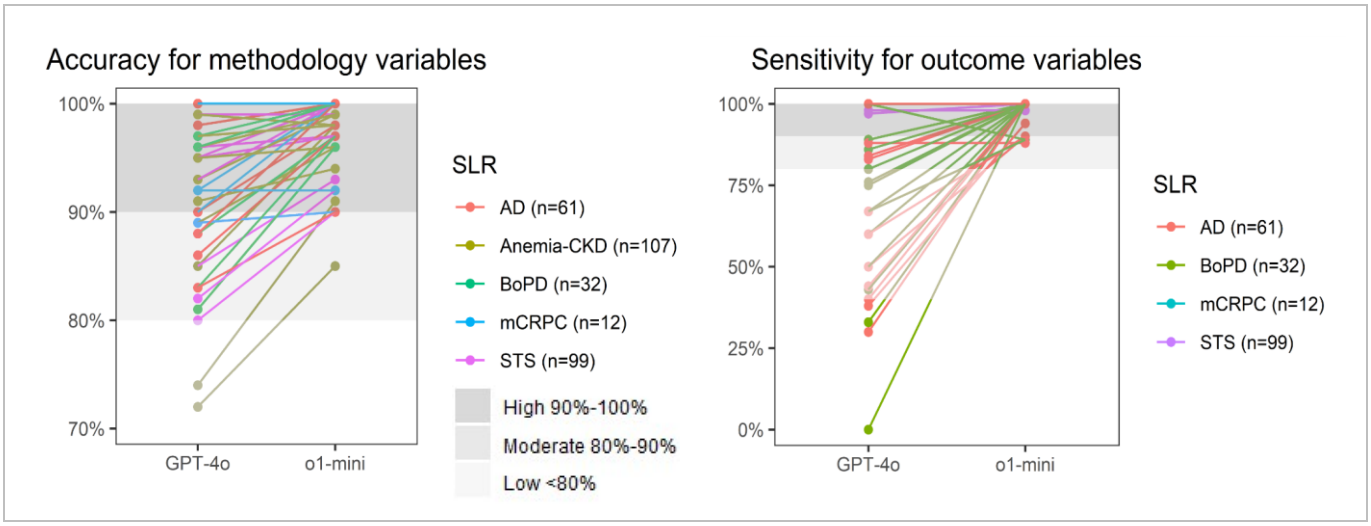
Comparison of o1-mini and GPT-4o models

Having previously tested GPT-4o using the same protocol, we compared those results with the results from o1-mini.

In general, although the GPT-4o model yielded reasonably strong results, the o1-mini model performed better, particularly with regards to outcome variables and sensitivity. The o1-mini model consistently achieved a tighter performance range, with **accuracy ranging from 85% and 100%** and **sensitivity from 88% to 100%**, indicating better reliability. In contrast, GPT-4o exhibited greater variability, with **accuracy ranging from 72% to 100%** and **sensitivity from 0% to 100%.**

- For all **methodology** variables, the accuracy either remained stable or improved with the o1-mini model (see **Figure 2**, which shows the performance of GPT-4o and o1-mini; each line represents a single variable; data from each review is shown in a different colour).
- Similarly, for all **outcome** variables, the sensitivity either remained stable or improved with the o1-mini model (**Figure 2**). The only exception was a single variable (mortality in the BoPD SLR), in which the sensitivity was lower with o1-mini compared to GPT-4o. In this case, for a single citation, both models identified the correct data (as evidenced by the corresponding rationale being very similar), but o1-mini classified it as 'No' (i.e., data on mortality not available), while GPT-4o correctly classified it as 'Yes'.

Figure 2: Accuracy for methodology variables and sensitivity for outcome variables using GPT-4o and o1-mini models across the SLRs



n=number of publications tested in each SLR project
AD=atopic dermatitis; BoPD=borderline personality disorder; CKD=chronic kidney disease; mCRPC=metastatic castration-resistant prostate cancer; STS=soft tissue sarcoma.

Notes:

1. Each point represents a variable (e.g., sample size, study design) for a single review; the line connects the AI accuracy and sensitivity values for that variable across the 2 models. On some occasions, the lines are overlapping due to the same values for a particular variable for both GPT-4o and o1-mini models across SLRs.
2. Each SLR is shown in a different colour
3. Number of methodology variables evaluated: STS and Anaemia-CKD (n=13 each), mCRPC (n=10), BoPD (n=8), and AD (n=7); Number of outcome variables evaluated: STS and mCRPC (n=4 each), BoPD (n=21), and AD (n=19).
4. For Anaemia-CKD, there were no relevant outcome variables for this analysis, and hence that SLR is not included in the figure for sensitivity.

Reflection and next steps

The o1-mini model achieved high performance across five SLRs in our testing of initial data extraction, with methodology variables averaging 97% accuracy (median 99%) and outcome variables averaging 99% sensitivity (median 100%). The accuracy for methodology variables ranged from 85–100% (only one variable fell below 90%), while sensitivity for outcome variables ranged from 88–100% with only 7 false negatives out of 2,214 extraction fields. This high level of consistency indicates that the model's performance is robust across SLRs investigating different types of parameters and heterogeneous variables across diverse disease areas and publication types (reviews and primary studies).

Accuracy and sensitivity were better with the newer AI model (o1-mini) than the earlier model (GPT-4o). The o1-mini model more effectively addressed the performance challenges associated with RW study–focused SLRs, which typically involve more heterogeneous and complex reporting.

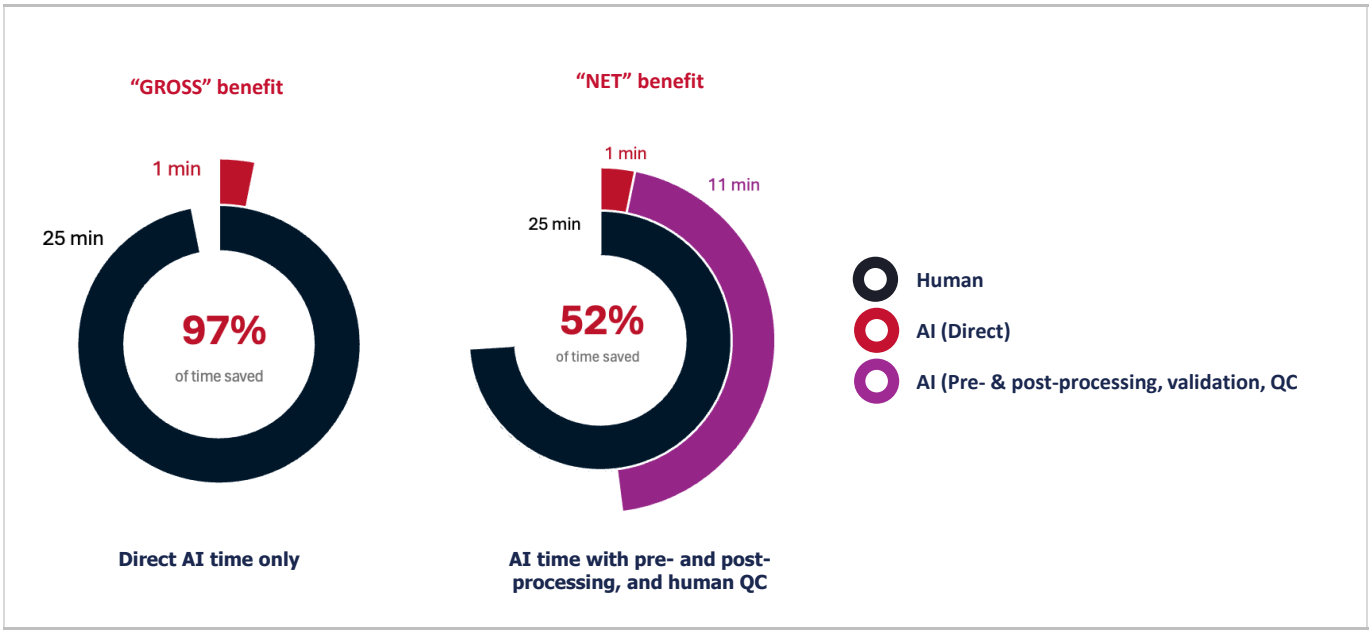
Based on these findings, we believe that AI can be used for creating a 'first-pass' initial extraction database, supplemented by human QC to ensure high-quality outputs (which is also consistent with the 'human-in-the-loop' approach recommended by the National institute for Health and Clinical Excellence [NICE][3] and other researchers in the field).[4,5,6]

Using our internal resourcing metrics, **Figure 3** summarises the time saved by using AI for initial extraction. When considering only the time taken by AI to complete initial extraction for one study, the gross time savings are 97%. In practice, one needs to account for the pre- and post-processing of data for the AI model, as well as the QC of the AI outputs by humans to ensure data quality. After accounting for this, net time saving is 52%, which still represents a substantial efficiency gain in this labour- and resource-intensive task.

Bridge is now offering this AI-enabled SLR component to clients alongside AI-enabled FTS (introduced in August 2024) and AI-enabled TiAB screening (introduced in December 2023).

We have also completed our validation work on both *full* data extraction and development of table narratives, and we will release white papers on each of these topics in the next few weeks.

Figure 3: Time benefit with AI vs Human implementation for initial data extraction from full texts



AI= artificial intelligence; QC=quality control

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. 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#:~:text=This%20position%20statement%20provides%20clarity,our%20guidance%20production%20is%20maintained. Accessed 11th February 2025.
4. 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.
5. 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.
6. 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.