

Artificial intelligence-assisted literature screening: a pilot study

Kim Wager,^a David Gothard,^a Andrew Liew,^b Eleanor J Raynsford^c

^aOxford PharmaGenesis, Oxford, UK; ^bOxford PharmaGenesis, Melbourne, VIC, Australia; ^cIpsen, London, UK

Presenter: Olivier Morteau, Ipsen US



For further information, please send your question(s) to Eleanor J Raynsford (eleanor.raynsford@ipsen.com).

To download the poster, please scan the Quick Response (QR) code.

Copies of this eposter obtained through the QR code are for personal use only and may not be reproduced without written permission from the authors.

Objective

- Manually screening medical literature for articles of interest is a substantial resource burden. Here we report the development and assessment of an artificial intelligence (AI)-assisted literature screening tool (**Panel 1**).

Research design and methods

- We evaluated a proprietary AI model for:
 - filtering literature search results for relevant articles, and
 - selecting the most relevant articles based on prespecified weighted prompts (study design, population, intervention etc.) (**Panel 1**).
- The model was developed and optimized using one manually screened data set (A; N = 82 articles) and tested on a second data set (B; N = 74 articles) (**Panel 1**).
- We compared AI with manual decisions, based on precision (avoiding false positives), recall (avoiding false negatives), and accuracy (proportion of matching decisions) (**Panel 1**).
- We also evaluated the model's ability to extract and summarize information (20 articles per data set), based on accuracy and likelihood of misinterpretation (Likert scale, 1 [most likely] to 4 [least likely]) (**Panels 1 and 4**).

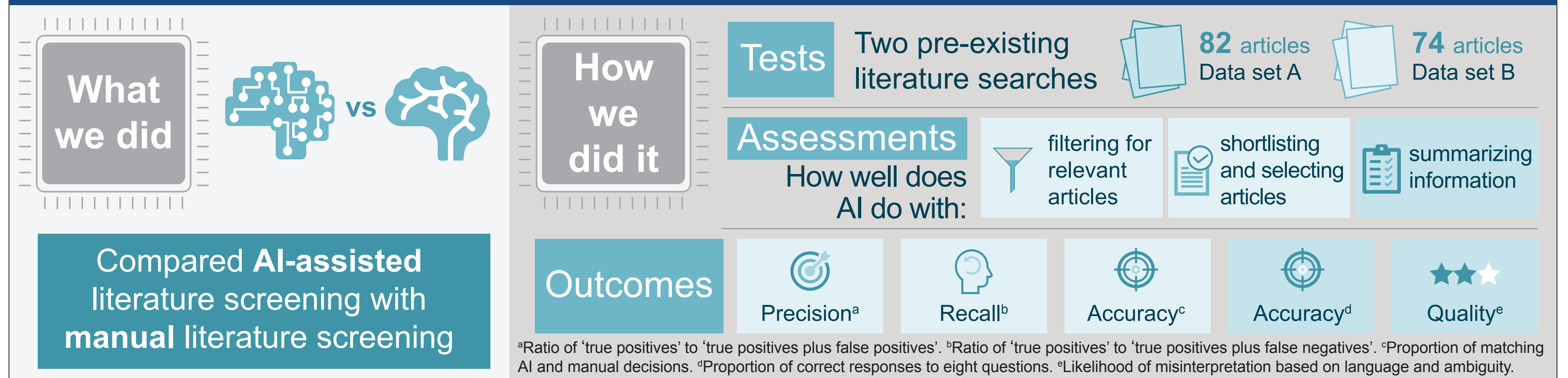
Results

- The model filtered for relevant articles with 78% (n = 64/82) and 73% (n = 54/74) accuracy in data sets A and B, respectively (inclusion precision, 78% [n = 46/59] and 75% [n = 39/52]; inclusion recall, 90% [n = 46/51] and 85% [n = 39/46]) (**Panel 2**).
 - The model missed only five and seven relevant articles in data sets A and B, respectively.
- However, the model performed suboptimally when selecting the most relevant articles (inclusion recall, 63% [n = 15/24] and 33% [n = 7/21]) (**Panel 3**).
- AI-generated summaries were of high quality; in both data sets, mean accuracy was 96%, and mean likelihood of misinterpretation was 3.6 (4 = lowest likelihood) (**Panel 4**).

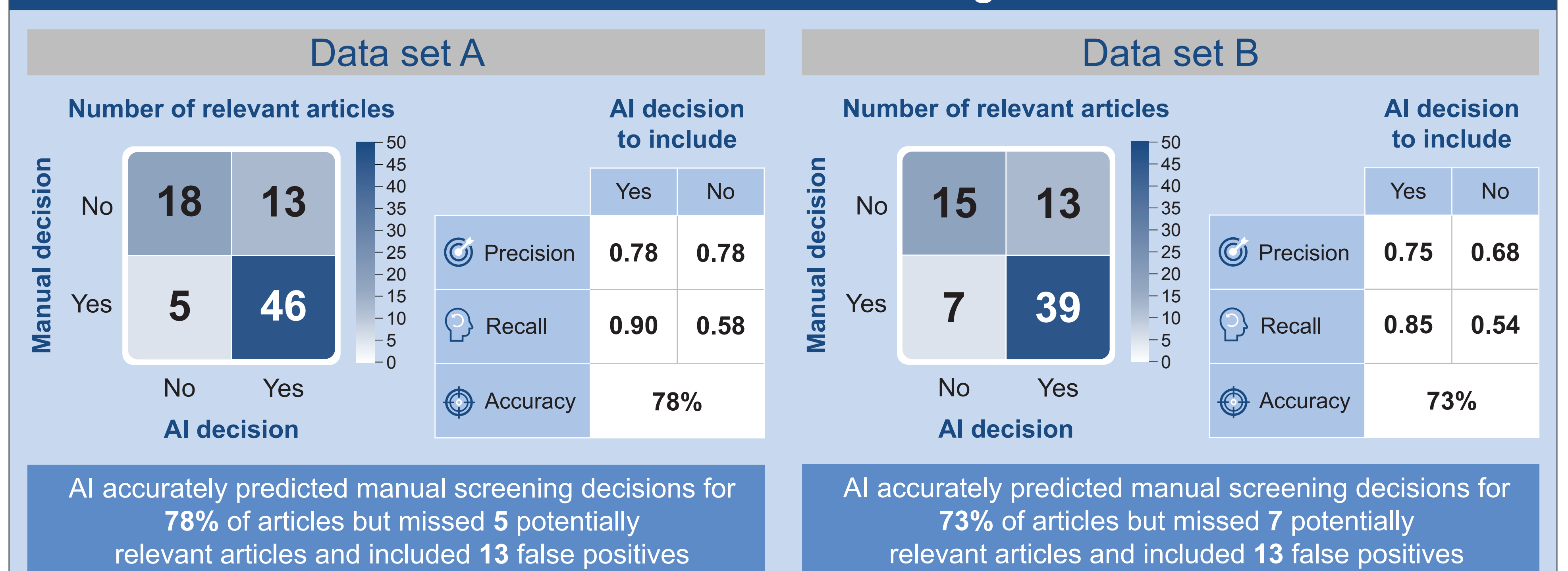
Conclusions

- We developed an AI model capable of extracting information and summarizing published articles.
- Although the model could accurately filter relevant articles, further work is needed to improve the model's ability to judge the level of relevance.
- At this stage, AI could be used to help to lower the workload burden by reducing the number of articles that require manual screening.

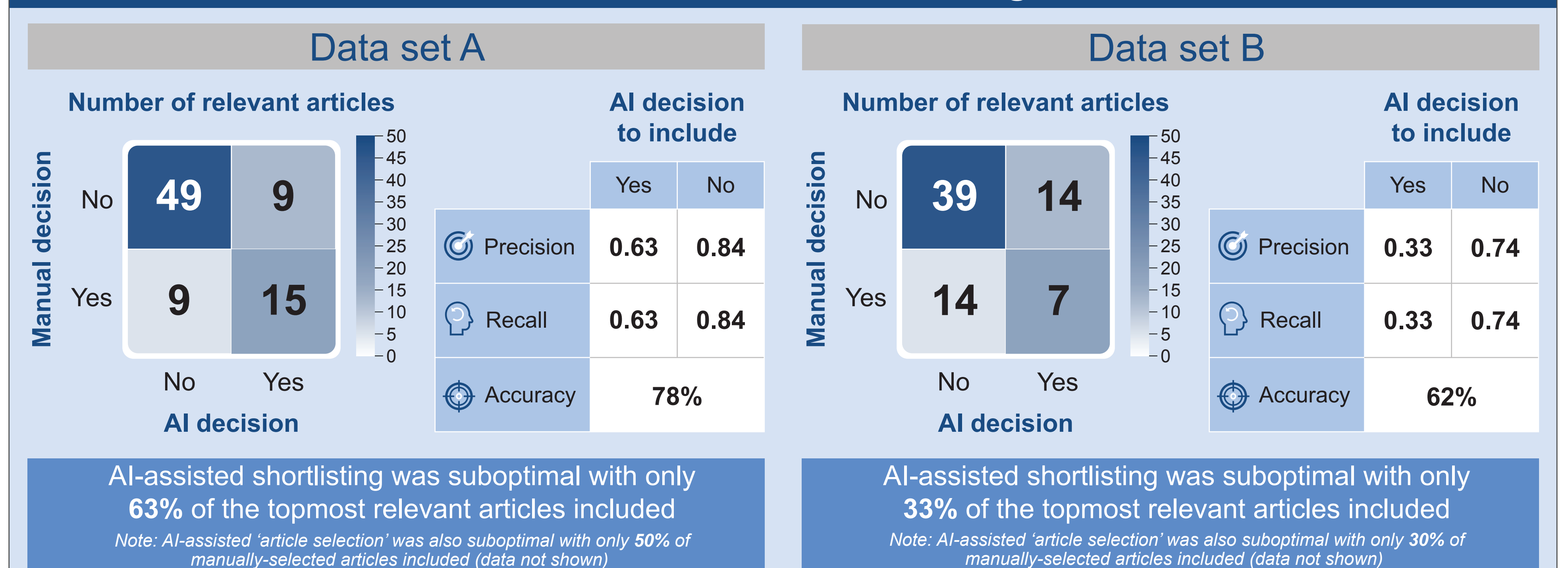
Panel 1: Objective and methods



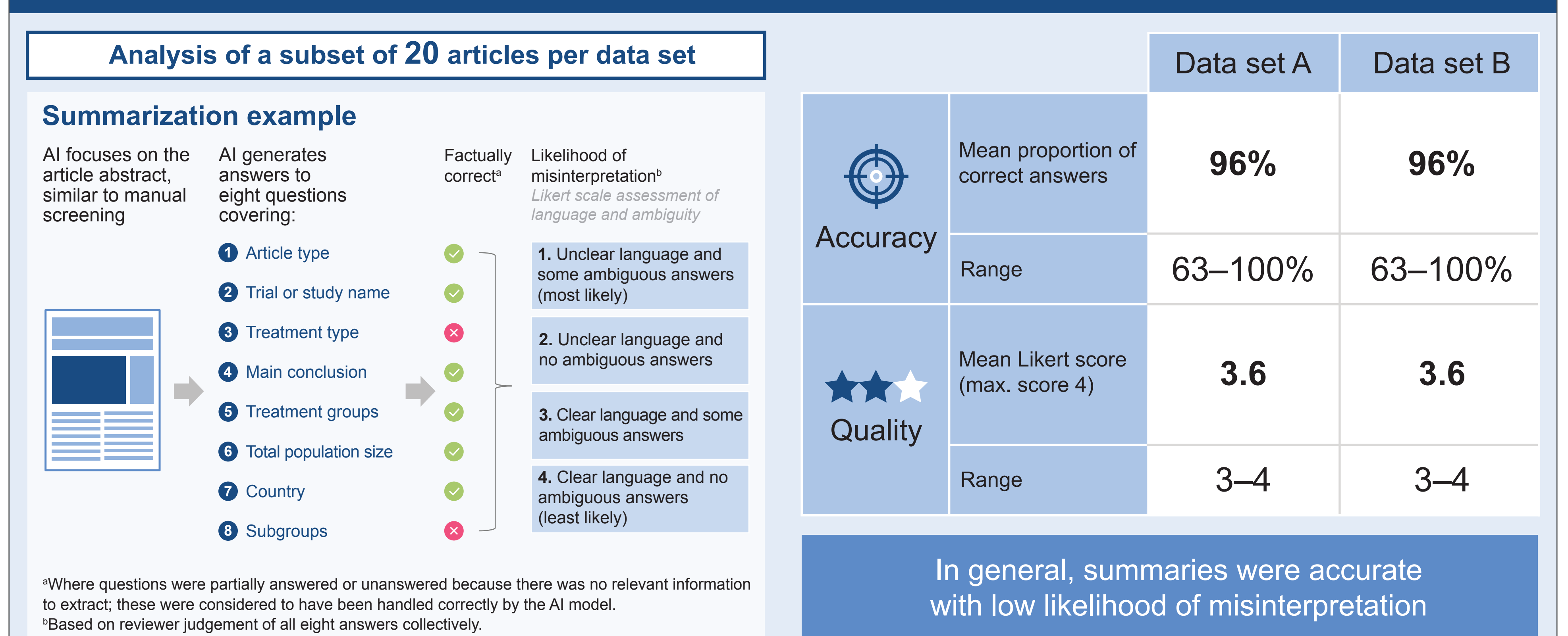
Panel 2: Article filtering



Panel 3: Article shortlisting



Panel 4: Article summarization



What we found

AI was able to identify articles of potential interest, but further work is needed to align AI decisions with manual decisions. AI was able to extract information and generate clear summaries, and could therefore be used to assist rather than replace manual screening.

Abbreviations AI, artificial intelligence.

Author contributions All authors provided substantial contributions to study conception/design or acquisition/analysis/interpretation of data; drafting of the publication or reviewing it critically for important intellectual content; and gave their final approval of the publication.

Disclosures KW, DG, AL: employees of Oxford PharmaGenesis. EJR: employee of Ipsen.

Medical writing support The authors thank Tamzin Gristwood, PhD, of Oxford PharmaGenesis, Oxford, UK, for providing medical writing support, which was industry sponsored in accordance with Good Publication Practice 2022 guidelines (GPP 2022).

Disclaimer Our research utilized a commercially available large-language model, but does not evaluate the model's function and performance; rather it evaluates our adapted research application.