

Concordance between generative artificial intelligence and human reviewers for screening of publications for a systematic literature review

Bernard Kerr,¹ Brian Norman,¹ Thierry Deltheil,¹ Valerie Moss,¹ Mary Coffey,² Iryna Shnitsar,² and Doreen Valentine²

¹Prime, London, UK; ²Bristol Myers Squibb, Princeton, NJ, USA

Explore research in more detail by scanning the QR code



Additional results and a video PLS can be accessed on the mobile-friendly website version of this poster.

Introduction

- Comprehensive literature searches are used to inform the development of systematic literature reviews (SLRs), publication plans, scientific communications platforms, and various other medical communications activities.
- The time, cost, and resource requirements to evaluate search results are barriers to undertaking literature searches.^{1,2}
- Generative artificial intelligence (genAI) tools may assist with literature searches, reducing associated costs and timelines; however, the accuracy and reliability of these tools require further investigation.
- This analysis investigated the agreement between a proprietary GPT4-powered literature review tool and validated human review for screening literature search results from a published SLR on analytical concordance of PD-L1 diagnostic assays by Prince EA et al.³
- The dataset was selected for this analysis because (1) the SLR was recently published and scientifically important; (2) it allowed the genAI model to be tested with challenging literature; and (3) it allowed testing of screening criteria with differing requirements.

Methods

- In total, 3477 unique publications were identified in literature searches and screened for inclusion in the published SLR by Prince EA et al³ (Figure 1).
 - 160 publications were available for genAI screening in the current analysis.
 - 97 publications that met both key inclusion criteria had screening data for human review of all criteria.
 - 63 open access publications that did not meet both key inclusion criteria were used as a control for the key inclusion criteria only.
 - The search methodology and publication selection are described in detail in the published SLR.³
- Question prompts were developed to screen the full text of all publications against 12 selection criteria from the published SLR methodology (6 inclusion and 6 exclusion criteria; Figure 1)³ using a proprietary GPT4-powered tool (prompts are shown in Online Table A).
 - GenAI responses were restricted to “yes”, “no”, or “unclear” (i.e., could not answer) for whether each criterion was met.
 - Three prespecified selection criteria from the original SLR were used to limit the original literature search and were not evaluated in this analysis.

Summary and Key Learnings

- This exploratory analysis highlights the potential for genAI-based tools to allow more efficient screening of literature search results than human review alone.
- Overall agreement between genAI and human review was high (>75%) for 8 selection criteria, moderate (69%) for 1 criterion, and poor (<50%) for 3 selection criteria.
 - Performance was unexpectedly high for the majority of selection criteria, given the dataset and selection criteria were chosen for their high complexity and variability.
- These results provide valuable insight into the strengths and limitations of genAI-based literature screening; however, comparison of this analysis with similar studies performed under different conditions should be done with caution.
- GenAI screening required <15 minutes of user time to run, supporting use of the tool to prioritize publications for human review, assist with categorizing publications, and reduce the risk of human error going undetected.

- This analysis offers various learnings for the future use of genAI in screening literature search results which are discussed in detail in the [online poster](#).
 - GenAI screening was most effective where information relevant to criteria was clearly stated in publications and closely matched the wording of the prompt.
 - Criteria with poor performance had prompts with a non-specific scope, related to complex concepts or technical information, or screened for information that was reported using highly variable language.
- Continued evolution of AI models, improved prompting approaches, and use of models with more relevant training data are expected to increase the effectiveness of AI screening.
- Further studies are planned to evaluate use of the genAI screening tool in other subject areas, and to investigate the reproducibility of gen AI responses.

Results

- GenAI screening was automated and required <15 minutes of total active user time to screen all 160 publications.
- Agreement statistics for inclusion criteria are shown in Table 1 and Figure 2.
- Overall agreement rates for inclusion criteria 1 and 2 were 88% and 86%, respectively (n = 160).
 - The overall agreement rate was 89% for meeting both key inclusion criteria in the overall dataset, 92% for the full-screening dataset that met both key inclusion criteria (n = 97), and 86% for the control dataset (n = 63) that did not meet both key inclusion by human review (Online Table B).
- Overall agreement rates were high (77%–100%) for inclusion criteria 3, 4, and 6 (n = 97); overall agreement for inclusion criterion 5 was 38%, with false-negative or unsure responses for 60% of publications.
 - Agreement statistics for congress abstract and journal publication subgroups were consistent with the overall dataset (data shown in Figure 2 in the online poster).
- The overall agreement rate for meeting all inclusion criteria was 42%; 65% of publications that met all inclusion criteria on human review had one or more false-negative result on AI screening.

Table 2. Agreement between genAI and human review for meeting individual exclusion criteria and ≥ 1 exclusion criteria

Agreement rate vs human, n/N (%)	Exclusion criteria						≥ 1 criterion
	1	2	3	4	5	6 ^a	
Overall agreement	84/97 (87)	67/97 (69)	94/97 (97)	45/97 (46)	48/97 (49)	97/97 (100)	45/97 (46)
False negative	–	1/2 (50)	1/1 (100)	0/40 (0)	1/1 (100)	–	0/43 (0)
False positive	8/97 (8)	26/95 (27)	2/96 (2)	52/57 (91)	46/96 (48)	0/97 (0)	52/54 (96)
Unsure	5/97 (3)	3/97 (3)	0	0	2/97 (2)	0	0

Yellow and orange shading indicate >25% and >50% disagreement with human, respectively. False-positive results reduce sensitivity and false-negative results reduce specificity. Denominators for false-positive and false-negative rates are the total number of publications that were negative and positive by human review, respectively, for that criterion. ^aExclusion criterion 6 was screened using two prompts that were combined using Boolean logic.

Figure 3. Agreement between genAI and human review for meeting individual exclusion criteria and ≥ 1 exclusion criteria

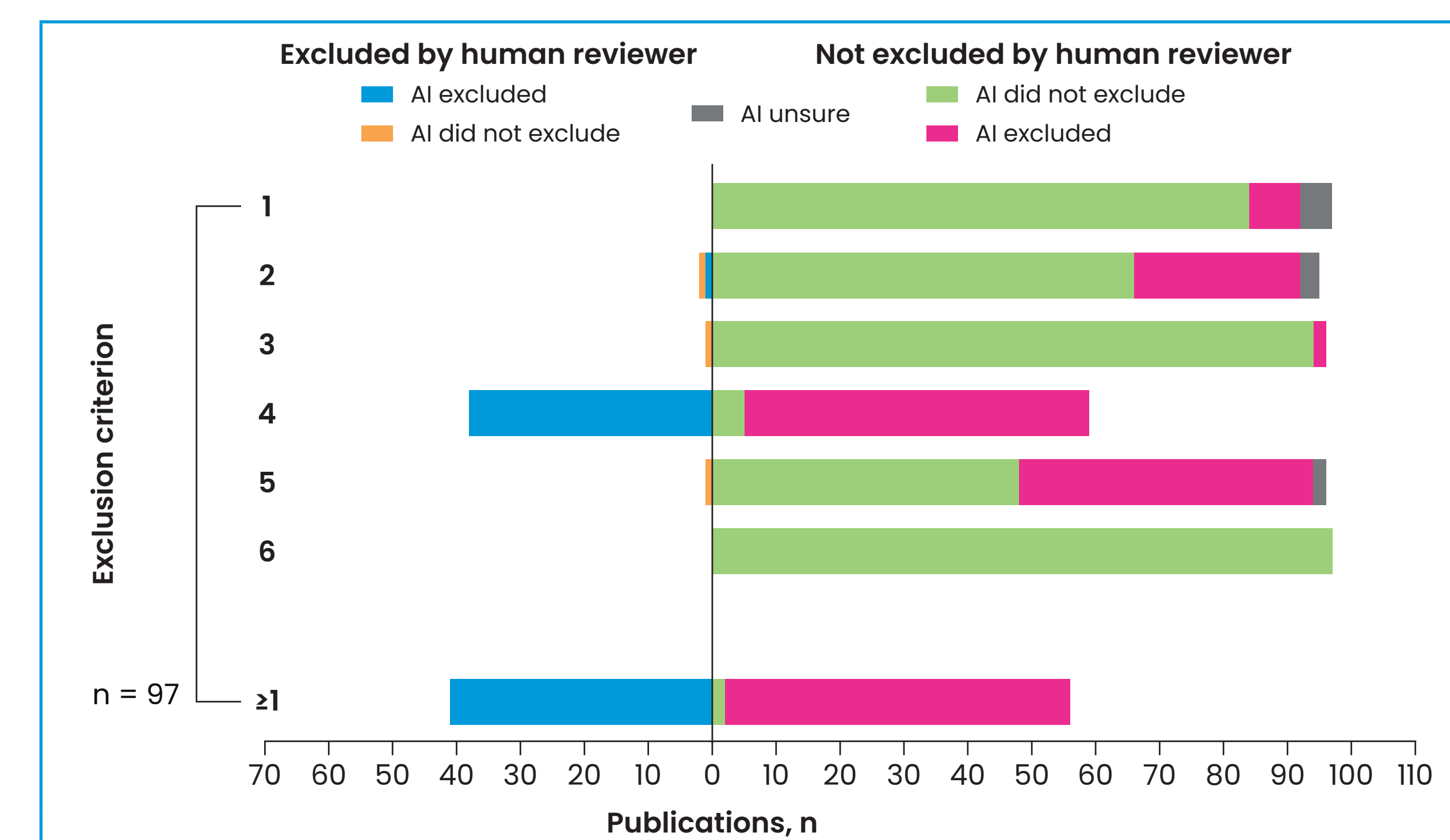
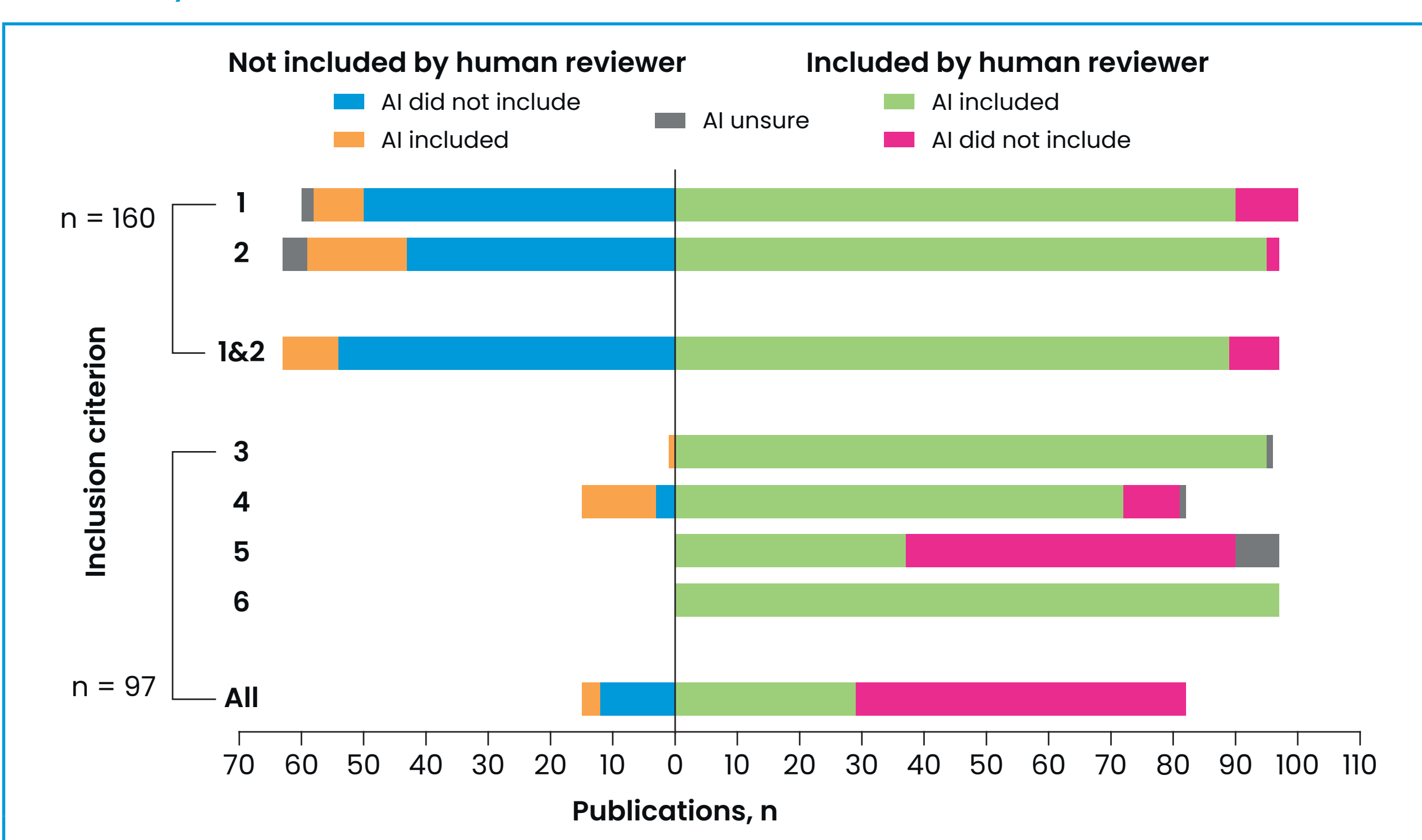


Table 1. Agreement between genAI and human review for meeting individual inclusion criteria, key inclusion criteria, and all inclusion criteria

Agreement rate vs human, n/N (%)	Inclusion criteria						Key criteria 1 & 2	All criteria
	1	2	3	4 ^a	5 ^a	6		
Overall agreement	140/160 (88)	138/160 (86)	95/97 (98)	75/97 (77)	37/97 (38)	97/97 (100)	143/160 (89)	41/97 (42)
False negative	10/100 (10)	2/97 (2)	0/96 (0)	9/82 (11)	53/97 (55)	0/97 (0)	8/97 (8)	53/82 (65)
False positive	8/60 (13)	16/63 (25)	1/1 (100)	12/15 (80)	–	–	9/63 (14)	3/15 (20)
Unsure	2/160 (1)	4/160 (3)	1/97 (1)	1/97 (1)	7/97 (7)	–	–	–

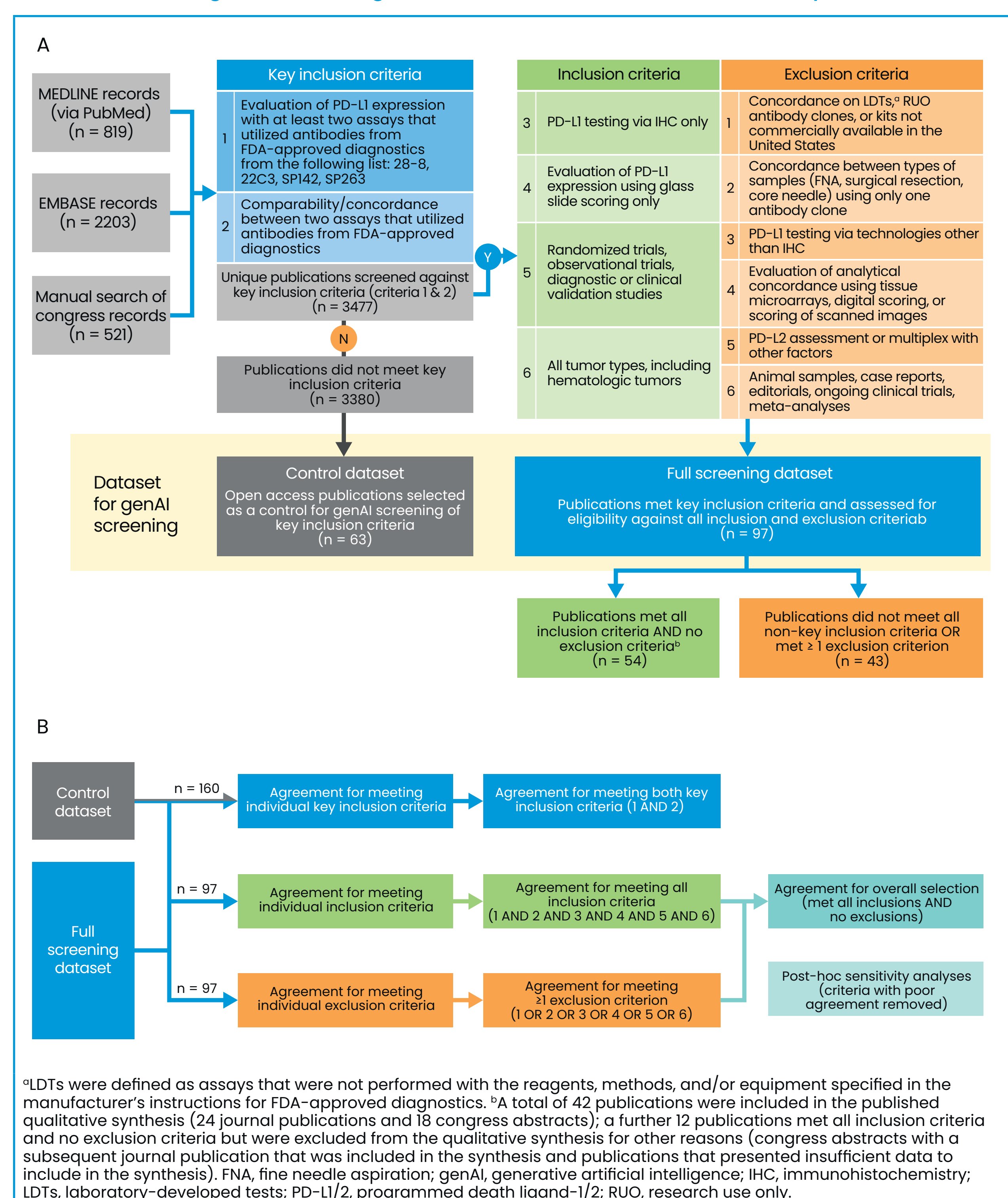
Yellow and orange shading indicate >25% and >50% disagreement with human review, respectively. False-negative results reduce sensitivity and false-positive results reduce specificity. Denominators for false-positive and false-negative rates are the total number of publications that were negative and positive by human review, respectively, for that criterion. ^aInclusion criteria 4 and 5 were screened using multiple prompts that were combined using Boolean logic.

Figure 2. Agreement between genAI and human review for meeting individual inclusion criteria, key inclusion criteria, and all inclusion criteria



- Agreement statistics for exclusion criteria are shown in Table 2 and Figure 3.
- Agreement between genAI and human review was variable across individual exclusion criteria (n = 97); overall agreement rates were high for criteria 1, 3, and 6 (87%–100%), moderate for criterion 2 (69%), and low for criteria 4 and 5 (44% and 49%, respectively).
 - The low number of publications excluded by human reviewers on most criteria skewed disagreement to false-positive results that incorrectly excluded publications (i.e., decreased sensitivity).
 - Agreement statistics for congress abstract and journal publication subgroups were consistent with the overall dataset (data shown in Figure 3 in the online poster).
- The overall agreement rate for meeting ≥ 1 exclusion criterion was low (46%); 96% of publications that did not meet any exclusion criteria on human review met one or more exclusion criteria on AI screening.

Figure 1. (A) Disposition of literature search results screened by Prince EA et al, composition of the dataset for genAI screening,³ selection criteria, and (B) flow of analyses



- Overall agreement, false positive, and false negative rates were calculated for genAI screening using human review as a gold standard.
- Associations of screening errors with word count and Flesch reading ease score of the AI-screened text files were investigated as surrogate measures of complexity.
- Post-hoc sensitivity analyses were performed to investigate the contributions of selection criteria with high disagreement to the overall screening outcomes by sequentially removing those criteria from the overall decision.
- To investigate the underlying reasons for incorrect screening results, additional prompts asked the genAI model to explain its decisions for inclusion criterion 1 and exclusion criterion 4.

- Performance of the genAI tool was poor for overall selection of publications based on all criteria (meeting all inclusion criteria AND no exclusion criteria).
 - All publications failed ≥ 1 selection criteria on AI screening (median 3; range, 1–7); 65 (67%) failed ≥ 1 inclusion criterion and 95 (98%) failed ≥ 1 exclusion criterion (Online Figure A).
 - 3 publications had screening results that agreed with human review on all criteria and were correctly excluded; the remaining 94 publications had a screening result that disagreed with human review on ≥ 1 criterion (median 2; range, 1–7) (Online Figure B).
 - Word count and readability were not associated with the number of criteria with disagreement (Online Figure C).
- The most permissive sensitivity analysis removed inclusion criterion 5 and exclusion criteria 2, 4, and 5 from the overall screening outcome for genAI (Online Figures D and E).
 - Removal of criteria with poor performance improved overall agreement with human review on matching criteria (73% for inclusion criteria and 89% for exclusion criteria), with the remaining disagreement suggesting an accumulation of screening errors across multiple criteria.
 - 39 (72%) of the 54 publications selected by human review on all criteria were selected by genAI in the sensitivity analysis, as were 31 of the 43 publications that were excluded by human review (false-positive rate 44%).
- Additional prompts asking the model to explain its responses for inclusion criterion 1 and exclusion criterion 4 identified a variety of errors that led to incorrect responses; findings are presented in detail in the online poster.

References

- Michelson M and Reuter K. *Contemp Clin Trials Commun*. 2019;16:100443.
- Borah R, et al. *BMJ Open*. 2017;7:e012545.
- Prince EA, et al. *JCO Precis Oncol*. 2021;5:953–973.

Disclosures

- Bernard Kerr, Brian Norman, Thierry Deltheil, and Valerie Moss are employees of Prime, London, UK.
- Iryna Shnitsar, Mary Coffey, and Doreen Valentine are employees of Bristol Myers Squibb, Princeton, NJ, USA.
- All opinions are the authors' own and do not necessarily reflect those of our employers.

Acknowledgements

- The Prime authors thank Bristol Myers Squibb for their support and collaboration, and for permission to use the Prince EA et al SLR in this analysis.
- We thank the Prime Editor, Digital, and Production teams for their support with development of this poster, and Jack Roweth, Prime, London, UK, for assistance with the analysis.