Poster Acceptance Code: MSR194

Qualitative Literature Reviews: A Comparison of Researcher and AI Screening of Articles to Inform Conceptual Model Development

Claire Burbridge[1], Lucy Lloyd-Price[1], Stacie Hudgens[2], Kristian Thorlund[3]

[1] Clinical Outcomes Solutions Inc., Folkestone, Kent, UK, [2]Clinical Outcomes Solutions Inc., Tucson, AZ, [3] BioSpark AI Technology Inc., Vancouver, BC, Canada

Background

- Artificial intelligence (AI) models are being used in systematic literature reviews, reducing researcher burden and improving efficiency. However, in structured (not fully systematic) literature reviews, such as those conducted in the Clinical Outcome Assessments (COA) space, aspects of the research such as study design, terminology and reporting are often not formal or standardized. It can therefore be challenging to develop focused yet comprehensive search strategies and screening criteria without compromising results.
- The broad scope of the search strategy often yields a high volume of returns, creating a burdensome workload for manual review and screening by researchers, or the need to limit the search (e.g., by date) to limit the number of results.
  - ➢ This is common in reviews to identify qualitative research and patient-focused insights exploring the patient lived experience.
- A novel large language AI model (LLM) **- COAScape AI -** is being developed and trained specifically to facilitate expert COA researchers screening literature for qualitative reviews.

Objectives

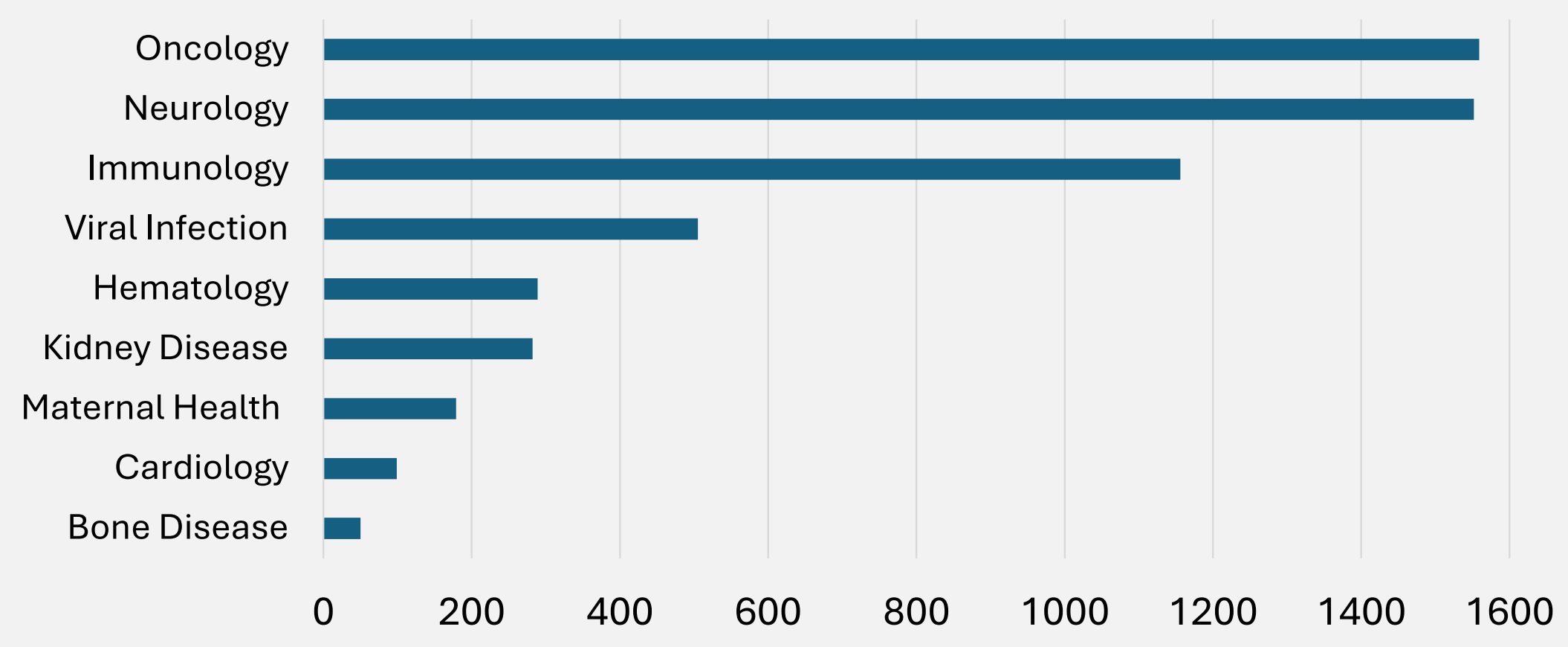
- To **compare title/abstract screening decisions** between researcher and the LLM when screening citations from qualitative literature review searches.
- To **explore areas of discrepancy** between researcher and the LLM to inform further application of the model.

Methods

Exploring Agreement in the Initial Test Datasets

- Data from 27 medical literature database reviews was used to compare researcher and the LLM title/abstract screening decisions.
- The reviews were previously conducted (in Ovid: Medline, Embase, PsycINFO) to identify qualitative research across therapeutic areas to inform conceptual models of the patient experience.
- The 27 qualitative reviews represented a range of conditions across key therapeutic areas (**see Figure 1**). This included a number of rare diseases.

Figure 1: Sum of Number of Citations by Therapeutic Area

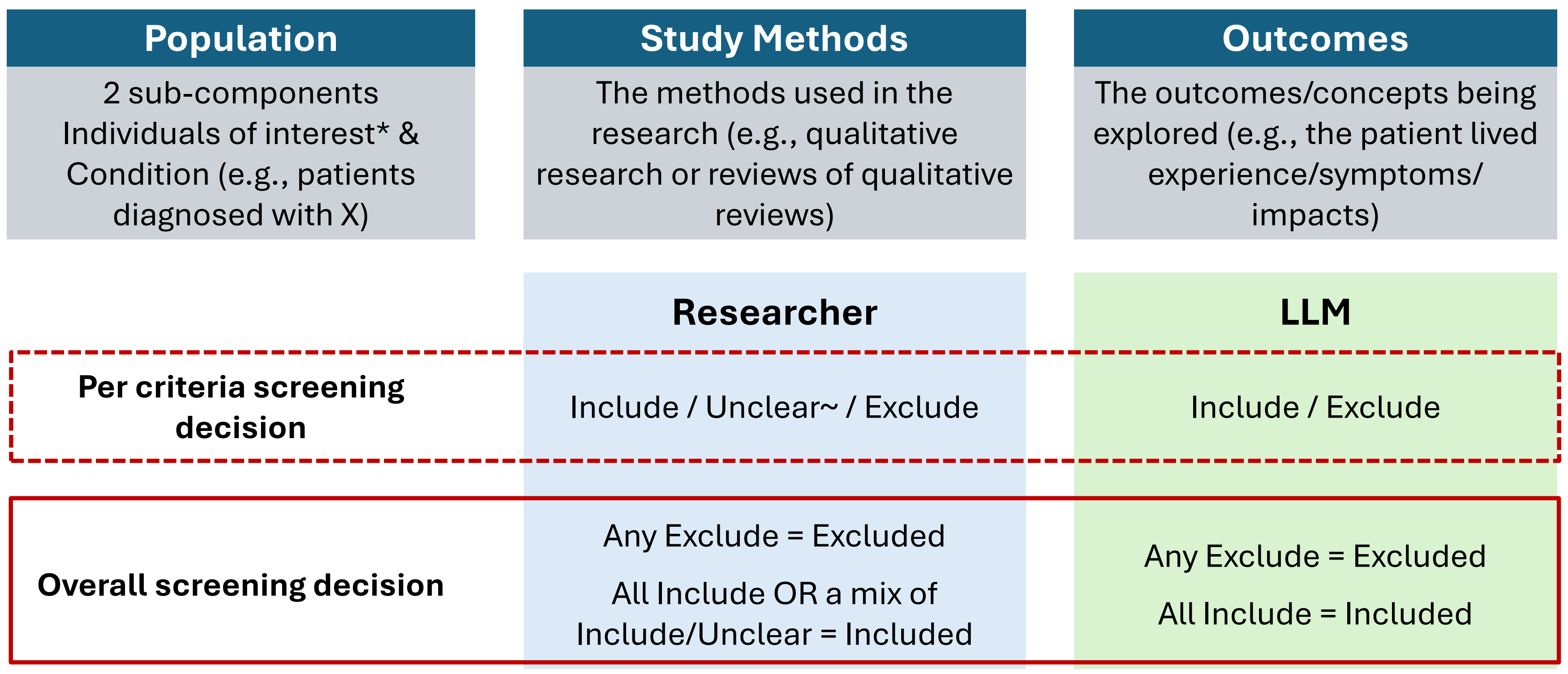


- Across all searches there were 5671 citations overall, (ranging from 20 to 942 per search); mostly in oncology (from 8 searches), neurology (from 7 searches), and immunology (from 5 searches).
- Screening decisions were annotated using researcher developed screening criteria based on PICO principles adapted for the context of a qualitative review (**Figure 2**).
  - ➢ Researcher developed criteria were used by both researcher and the LLM.
  - ➢ Screening decisions were compared **per criteria** as well as the **overall decision** for include (relevant) or exclude (not relevant) for each citation.
  - ➢ The LLM used 75% data within each dataset to learn, and 25% to test and screen.

Detailed Exploration of Agreement

- A sub-sample of the screened datasets (n=9) was reviewed in greater depth to explore screening discrepancies between first researcher and the LLM.

Figure 2: Researcher Developed Screening Criteria and Researcher/LLM Screening Decision Making



Population
2 sub-components Individuals of interest* & Condition (e.g., patients diagnosed with X)

Study Methods
The methods used in the research (e.g., qualitative research or reviews of qualitative reviews)

Outcomes
The outcomes/concepts being explored (e.g., the patient lived experience/symptoms/ impacts)

	Researcher	LLM
Per criteria screening decision	Include / Unclear~ / Exclude	Include / Exclude
Overall screening decision	Any Exclude = Excluded All Include OR a mix of Include/Unclear = Included	Any Exclude = Excluded All Include = Included

*Note: individual of interest may/may not be the research participant (e.g., patient/caregiver reporting on patient experience)
~'Unclear' was applied in cases where the researcher felt they needed to discuss with senior colleagues or consult full-text for more information to make a definite decision on relevance.

Results

- Across the 27 datasets, the level of agreement between the researcher and LLM screening decisions was high at 86.5% (**Figure 3**). For all but 4 searches, agreement was 75% or above on all screening criteria (**Figure 4**).

Figure 3: Average Level of Agreement (%) on Screening Decisions

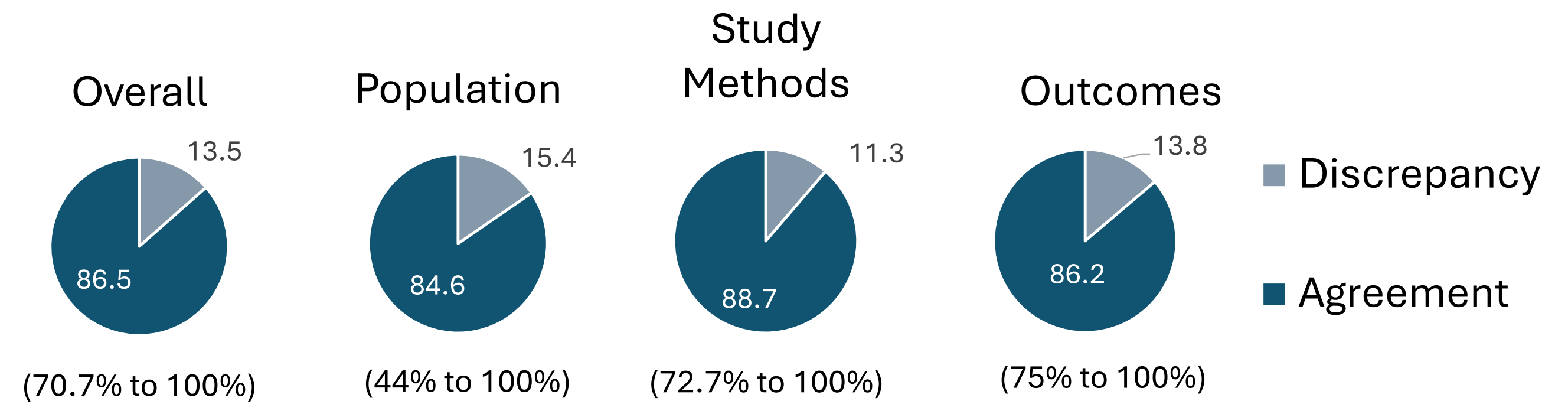
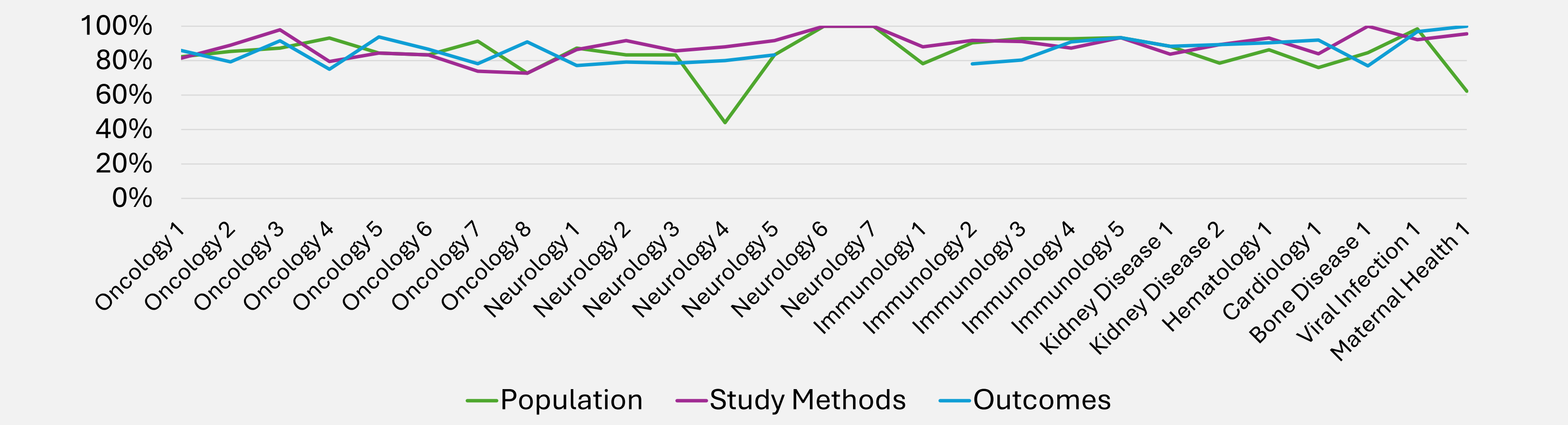


Figure 4: Level of Agreement (%) on Screening Decisions By Dataset



- Across the 9 datasets, screening decisions were compared for 25% citations (test dataset; 401 citations; 1,203 decision pairs across the 3 criteria).
- **The LLM correctly predicted 1,041 (86.5%) of these decisions;** only 162 (13.5%) were discrepant, meaning the first researcher and the LLM did not align (**Table 1**).
  - ➢ Of these discrepancies, 53 (32%) were attributed to the LLM and 109 (67%) to the researcher.
  - ➢ Almost half (N=49; 45%) of the discrepancies attributed to the researcher were instances where they had indicated 'unclear' for a screening criteria.
  - ➢ Upon senior review, it was possible to determine a firm decision for include/exclude against all of those that were 'unclear', without reference to full-text; 37 decisions upon senior review aligned with LLM; 12 did not).
- Following senior review of the 'unclear' ratings, only 125 of the 1,203 decisions remained discrepant **increasing the level of agreement to 89.6%.**

Table 1: Number of Discrepancies Per Screening Criteria Attributed To LLM Error or Researcher

Dataset	Total # citations	# citations compared	Total # discrepancies	Attributed to LLM				Attributed to Researcher				*Unclear Rating (subgroup of those attributed to researcher)*			
				Population	Study Method	Outcomes	TOTAL	Population	Study Method	Outcomes	TOTAL	*Population*	*Study Method*	*Outcomes*	*TOTAL*
Kidney Disease 2	111	**28**	**14**	2	2	2	**6**	5	1	2	**8**	*0*	*0*	*0*	***0***
Oncology 3	185	**47**	**21**	4	1	1	**6**	3	7	5	**15**	*1*	*6*	*2*	***9***
Kidney Disease 1	171	**43**	**26**	4	1	1	**6**	1	10	9	**20**	*1*	*6*	*5*	***12***
Oncology 4	176	**44**	**25**	1	7	6	**14**	2	3	6	**11**	*0*	*2*	*3*	***5***
Oncology 5	125	**32**	**20**	2	1	1	**4**	3	9	4	**16**	*1*	*5*	*3*	***9***
Maternal Health 1	179	**45**	**19**	9	0	0	**9**	8	2	0	**10**	*0*	*0*	*0*	***0***
Oncology 6	119	**30**	**14**	4	0	0	**4**	1	5	4	**10**	*1*	*2*	*2*	***5***
Viral Infection 1	505	**127**	**23**	0	3	1	**4**	2	14	3	**19**	*1*	*6*	*2*	***9***
Neurology	20	**5**	**0**	0	0	0	**0**	0	0	0	**0**	*0*	*0*	*0*	***0***
TOTAL	**1591**	**401**	**162 (13.5%)**	**26**	**15**	**12**	**53**	**23**	**51**	**33**	**109**	***5***	***27***	***17***	***49***

- Other reasons for the researcher attributed discrepancies were mainly due to complexity of the condition of interest or research involving multiple conditions or individuals of interest.
- Most common reasons for the LLM attributed discrepancies were citations in which terms related to qualitative research were referred to in an abstract when they were not the core study methods.

Conclusions

- There is a **high level of agreement** between expert researcher and the LLM in title/abstract screening.
- The level of agreement increased following senior review of the initial researcher decisions. This demonstrates that researcher fatigue can occur when manually screening large datasets, which plays a key part in screening decision inconsistency.
- Findings highlight **the potential of AI (LLM) to facilitate the researcher** in efficient screening of qualitative literature reviews that are being used to inform the development of conceptual models of the patient experience in the context of COA research.
- Use of the LLM as an initial screening tool supports faster screening decisions and the ability to screen larger datasets without compromising on quality.
- Since this initial test dataset review, additional datasets have been used to **further train the LLM** and the level of agreement has increased (data to be published).

COA AI

This study was sponsored by COA-AI | Editorial and programming support was provided by COA-AI

Poster presented at ISPOR Europe 2024
17th to 20th November 2024
Barcelona International Convention Center, Barcelona, Spain
Poster Session 5

Acknowledgements
We would like to thank other members of the COS and BioSpark teams who have contributed to the reviews and training of the AI model.
Special thanks to Angie Viada, Reza Jafar, Elliott Blatt, Cindy Umanzor-Figueroa, and Melanie Hughes.