THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Collaborative Approach to Meta Analysis and ·C·A·M·A·R·A·D·E·S·
Review of Animal Data from Experimental Studies

Evaluation of Elicit.AI in replicating three systematic reviews

Dr Charis Wong
Honorary Senior Clinical Lecturer, University of Edinburgh
Consultant Neurologist, NHS Fife
charis.wong@ed.ac.uk

Co-authors: Harris Khalid, Florence Do, Kat Stefanska, Mohammed Abomostafa, Gemma Dalgety, Professor Malcolm Macleod

NIHR Evidence Synthesis Group | 16th July 2026

edin.ac/incr

Aim

- To evaluate the performance of three AI-enabled systematic review platforms in replicating three recently published systematic reviews

- Protocol pre-registered on OSF 1st July 2025 (https://osf.io/8nf6h/files/qadcw)

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Methods: Evaluation overview

- Random selection of systematic reviews (SR) published in past year by CAMARADES, JBI, Cochrane containing quantitative analysis

- Systematic review platforms for evaluation: Elicit.AI, Nested Knowledge, Scispace

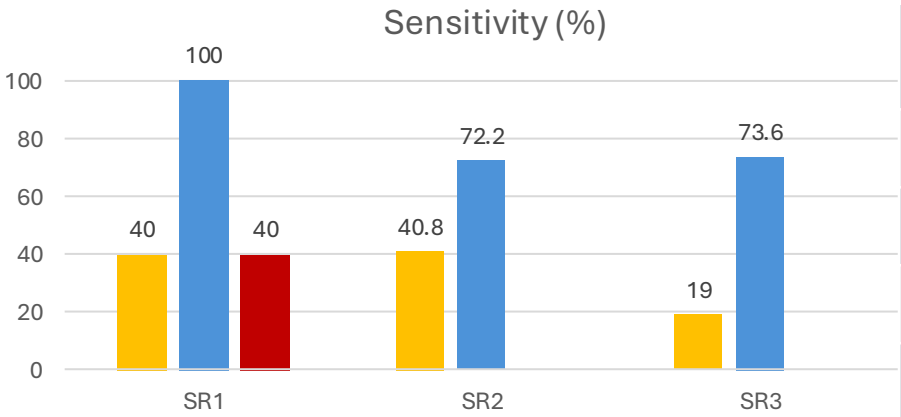
- Evaluation of these platforms on each SR in a 3x3 design across systematic review stages:
  - Search/Screen
  - Annotation (PICO, Risk of bias)
  - Quantitative analysis

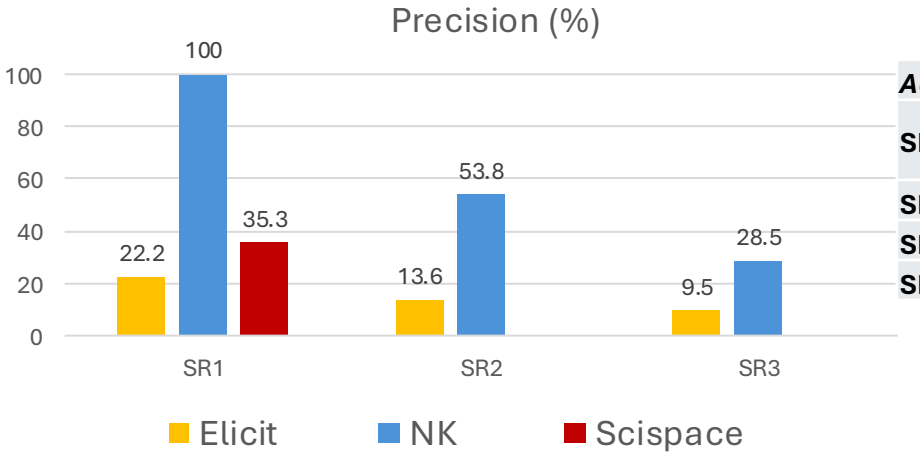
	SR1	SR2	SR3
SRP1	Evaluation 1	Evaluation 4	Evaluation 7
SRP2	Evaluation 2	Evaluation 5	Evaluation 8
SRP3	Evaluation 3	Evaluation 6	Evaluation 9

SR overview

N Citations	**SR1: CAMARADES** Effect of trace amine-associated receptor 1 (TAAR1) agonism for psychosis in animals	**SR2: JBI** Comparison of diagnostic accuracy of rapid antigen tests for COVID-19 compared to the viral genetic test in adults	**SR3: Cochrane** Community care navigation intervention for people who are at risk of unplanned hospital presentations
Retrieved	1849	3122	7461
TiAb included	20	580	181
FT included	15	261	19
Passed Critical appraisal		143	
Included in data extraction	15	91	19 (10 in quantitative)
Notes		Details for excluded TiAb unavailable	Details for excluded TiAb unavailable 'Table 1' less typical

Search and screen





Breakdown of original included publications by status on Elicit review

Elicit label	SR1	SR2	SR3
Include	6	58	4
Exclude	0	0	0
Not retrieved	9	84	17

Additional included studies by Elicit (unique in parentheses)

SR	Published before SR		Published after SR	
	Appropriate	Not appropriate	Appropriate	Not appropriate
SR1	2	15	1	3
SR2	70 (50)	121 (116)	64	23
SR3	3	25	1	1

Annotations

Percentage of tags in original review accurately tagged by AI features in SRPs						
Evaluation	**Population**	**Intervention**	**Comparison**	**Outcome**	**ROB**	**Overall**
SR1-Elicit	60.0%	55.6%	66.7%	100.0%[A]	51.7%	57.4%
SR2-Elicit	NA	85.5%	60.3%	68.9%	66.2%	68.8%
SR3-Elicit	50.0%	66.7%	100.0%	66.7%	33.3%	50.0%
Overall Elicit	55.0%	69.2%	75.7%	78.5%	50.4%	58.7%

Percentage of publications given additional tags by AI features in SRP (percentage appropriate in parenthesis)						
Evaluation	**Population**	**Intervention**	**Comparison**	**Outcome**	**ROB**	**Overall**
SR1-Elicit	80.0% (0%)	55.6% (20%)	66.7% (0%)	100.0% (33%)	0%	26.6% (12%)
SR2-Elicit	NA	6.0% (0%)	0.00%	13.7% (48%)	33.8% (0.5%)	24.0% (6%)
SR3-Elicit	50.0% (100%)	66.7% (100%)	0.00%	66.7% (100%)	66.7% (0%)	45.5% (43.3%)

[A] No quantitative outcomes evaluated. Elicit cannot extract data from figures (July 202)

Quantitative analysis

“At this time, Elicit does not perform calculations on data that is extracted from papers. It can extract the data for you, but it won't manipulate it further. You would need to export the data from Elicit and perform those calculations externally.”

For animal and preclinical studi on behavioural measures relevar ments of psychosis? What are th preclinical animal experiments o agonists on neurobiological meas ergic, glutamatergic and seroton ments of psychosis? Which are th effects? If a causal pathway (or p findings of the aforementioned r living systematic review, is there hypothesis?

TAAR1 agonists consistently reduce psychosis-r social interaction impairments, and cognitive dy models without inducing catalepsy or extrapy midbrain dopaminergic neuron firing, reduction activity in the dorsal raphe, and engagement Akt/GSK3β/BDNF signaling pathways—a causal electrophysiological evidence demonstrating tha presynaptic modulation.

Abstract

TAAR1 agonists consistently reduce psychosis-rel pharmacological models using NMDA receptor a attenuate hyperlocomotion, improve prepulse inl in novel object recognition tasks. Similar efficacy KO), where multiple TAAR1 agonists dose-depen duces PCP-induced but not amphetamine-induce matergic dysfunction. Critically, TAAR1 agonists at doses up to 100 mg/kg for SEP-856 and no extr for compound 50B.

Mechanistically, TAAR1 agonists modulate all th inhibit dopaminergic neuron firing in the ventral striatal dopamine synthesis capacity by 44-50% a bition of dorsal raphe nucleus firing and modulat is evidenced by reversal of NMDA antagonist-ind level, TAAR1 activation elevates cAMP and indu Akt/GSK3β/BDNF pathway. Direct evidence sup that inhibition of midbrain dopamine neurons c heterodimer formation that modulates dopamine

What is the diagnostic accuracy of the curren of-care rapid antigen tests relative to viral g for the diagnosis of COVID-19/SARS-CoV-2

Rapid antigen tests for SARS-CoV-2 achieve 80-100% sensitivity within the first week of symptom onset or in individuals with h 22-65% in asymptomatic individuals or those tested beyond one w identifying infectious cases during the acute phase but requiring or asymptomatic presentations.

Abstract

This review synthesized evidence from 80 studies evaluating rapid in adults compared to RT-PCR. Overall sensitivity ranged from 27.5 ceeded 97% in most studies [3–17]. This apparent heterogeneity re rather than test unreliability. In symptomatic individuals within 5 sensitivity [3, 4, 14, 15, 18, 19], increasing to 92-100% within the loads (Ct <25), sensitivity exceeded 90% regardless of symptom at Ct 25-30 [6, 7, 21, 28] and below 30% at Ct >30 [7, 15, 21, 27, markedly lower at 22.6-63.6% [16, 31–35], though this improved to [16, 32].

Ag-RDTs reliably identify individuals during the period of peak week of symptoms, Ct <25-30), achieving sensitivities of 89-98% ficity [3, 4, 13–17]. Performance declines after the first week [15, 30, 40], limiting utility as a rule-out test in these contexts withou comparable accuracy to professional testing (82.5% vs 85.0%) [41 oropharyngeal approaches [23]. Studies evaluating different SA no substantial variant-specific differences in performance when a

Primary: to assess the effects of care navigation, delivered in the community, on hospital presentations and patient-reported outcome measures in people at risk of unplanned hospital presentations. Secondary: to assess whether the effects of community care navigation differ according to the type of clinician delivering the intervention and the populations receiving the intervention.

Community-based care navigation interventions inconsistently reduce hospital utilization but more reliably improve patient-reported outcomes such as self-management and care experience, with benefits most evident in older adults and populations facing language or social barriers, and moderated more by intervention intensity, fidelity, and targeting than by the specific type of clinician delivering the intervention.

Abstract

Community-based care navigation interventions show mixed effects on hospital utilization among populations at risk of unplanned presentations. Twelve of 27 studies reporting outcomes demonstrated statistically significant reductions in at least one measure of hospital utilization, including rehospitalization rates, length of stay, or emergency department visits. Notable positive findings include a borderline-significant reduction in all-cause rehospitalization with a 12-month cardiac care navigator program (62.6% vs. 66.4%, p = 0.05) [1], significant reductions in hospital days with a standardized community health worker intervention (−172.3 days per 1,000 members per month, p = 0.022) [2], and a 21% reduction in unplanned readmissions through pharmacist-GP collaboration [3]. However, several large trials found no significant effects on hospital utilization, including a nurse-led multifactorial care program (n = 2,283) [4] and a transitional care program for the highest-risk patients in Singapore (n = 840) [5]. Patient-reported outcomes, particularly self-management, patient experience, and perceived quality of care, improved in multiple trials even when hospital utilization was unaffected [5–8].

Effects differ by population and clinician type, though direct comparisons are limited. Patient age is a consistent moderator: navigation reduced hospital utilization among patients aged ≥60 years (4.1% absolute reduction in 30-day readmission, p < 0.05) but paradoxically increased it among younger patients (11.8% absolute increase, p < 0.05) [9], with benefits for older patients sustained over 180 days [10]. Populations facing language barriers also showed disproportionate benefit (p = 0.03) [1]. Regarding clinician type, a three-arm trial found that GP-led comprehensive geriatric assessment applied at clinical discretion significantly reduced unplanned admissions (aOR 0.57, p = 0.020), whereas systematic nurse-led assessment did not (aOR 0.81, p = 0.31) [11]. Across studies, intervention structure, fidelity, and targeting mechanisms appear to matter more than the specific professional background of the navigator, with multifaceted, higher-fidelity interventions directed at moderately high-risk populations in settings with suboptimal baseline coordination most consistently demonstrating benefit.

Conclusions comparison: SR1

Elicit

TAAR1 agonists consistently reduce psychosis-relevant behaviors (hyperlocomotion, prepulse inhibition deficits, social interaction impairments, and cognitive dysfunction) across NMDA antagonist and dopamine hyperactivity models without inducing catalepsy or extrapyramidal symptoms.

Original

In animal studies, TAAR1 agonists improved hyperlocomotion compared to control (N=13 studies, k=41 experiments, SMD=1.01, 95%CI: 0.74, 1.27), but seemed less efficacious compared to dopamine D2 receptor antagonists (N=4, k=7, SMD=-0.62, 95%CI: -1.32, 0.08).

The data for prepulse inhibition impairment, the second primary outcome, came from just one study in animal models of psychosis, so it was not analysed further.

There were limited data for secondary efficacy outcomes, with 4 studies reporting data for cognitive function. There were no clear differences between TAAR1 agonists and control (N=4, k=5, SMD=0.80, 95% CI: -0.30, 1.90)

Combined publication of human and animal data

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Conclusions comparison: SR2

Elicit

Rapid antigen tests for SARS-CoV-2 achieve 80-100% sensitivity and >97% specificity in symptomatic adults tested within the first week of symptom onset or in individuals with high viral loads (Ct <25), but sensitivity declines to 22-65% in asymptomatic individuals or those tested beyond one week of symptoms, making them highly accurate for identifying infectious cases during the acute phase but requiring RT-PCR confirmation when negative in later-stage or asymptomatic presentations.

Original

RATs can identify individuals who have COVID-19 with high reliability (positive predictive value 97.7%; negative predictive value 95.2%) when considering overall performance. However, the lower level of sensitivity (67.1%) suggests that negative test results likely need to be retested through an additional method.

Conclusions: Most reported RAT brands had only a few studies comparing their performance with RT-PCR. Overall, a positive RAT result is an excellent predictor of a positive diagnosis of COVID-19. We recommend that Roche's SARS-CoV-2 Rapid Antigen Test and Abbott's BinaxNOW tests be used in primary care settings, with the understanding that negative results need to be confirmed through RT-PCR…

Conclusions comparison: SR3

Elicit

Community-based care navigation interventions inconsistently reduce hospital utilization but more reliably improve patient-reported outcomes such as self-management and care experience, with benefits most evident in older adults and populations facing language or social barriers, and moderated more by intervention intensity, fidelity, and targeting than by the specific type of clinician delivering the intervention.

Original

Community care navigation for people at risk of unplanned hospital presentations is likely to reduce hospital admission rates within 12 months (365 days) and increase outpatient appointments within one month (30 days) compared to usual care, with moderate to high certainty of evidence. Results showed little to no effect on hospital admissions within one month (30 days) or on emergency department presentation rates compared to usual care. The evidence is very uncertain about the effect of community care navigation on health-related quality of life and quality of care.

Conclusion

- Variation in performance between SRs noting variation in size, scope, data type (e.g. figures)
- Responsible use of AI in evidence synthesis (RAISE) 2026 guidance

  "When considering using an AI system or tool, be critical of its evaluations to understand whether it does what it states at an adequate level, as well as its limitations, and whether it can be applied to the context of the specific synthesis."

- Results limited to version evaluated; quickly superseded
  - *Need better methods for more efficient and timely evaluation of emerging AI evidence synthesis tools and platforms*

Acknowledgements

Harris Khalid, Florence Do, Kat Stefanska, Mohammed Abomostafa, Gemma Dalgety, Malcolm Macleod

Additional reviewers for screening SR3-Nested Knowledge:

Leonor Rodriguez, Matheus Gallas-Lopez, Stephen Malden, Kaitlyn Hair, Gillian Currie, Yvonne Khor, Nadia Soliman, Khrystyna Stoska

✉ charis.wong@ed.ac.uk

References

SR1: Siafis S et. al. Trace amine-associated receptor 1 (TAAR1) agonism for psychosis: a living systematic review and meta-analysis of human and non-human data. Wellcome Open Res. 2024 Apr 11;9:182. doi: 10.12688/wellcomeopenres.21302.1.

SR2: Hirabayashi E et. al. Comparison of diagnostic accuracy of rapid antigen tests for COVID-19 compared to the viral genetic test in adults: a systematic review and meta-analysis. JBI Evid Synth. 2024 Oct 1;22(10):1939-2002. doi: 10.11124/JBIES-23-00291.

SR3: Pang RK, Shannon B, Collyer T, Srikanth V, Andrew NE. Community care navigation intervention for people who are at risk of unplanned hospital presentations. Cochrane Database of Systematic Reviews 2025, Issue 6. Art. No.: CD014713. DOI: 10.1002/14651858.CD014713.pub2.

RAISE guidance (Responsible use of AI in evidence synthesis): https://osf.io/fwaud/overview