Evaluating paper screening capabilities of an AI assistant for a systematic review of HLA molecular mismatching and kidney transplant outcomes

Jacob Ujhazy[1], Dhruva Nilakantan[2], Maria P. Bettinotti[3], Chethan Puttarajappa[1], Vikas Dharnidharka[2], Loren Gragert[4]

[1]University of Pittsburgh Medical Center, Pittsburgh, PA, United States; [2]Rutgers University Robert Wood Johnson Medical School, New Brunswick, NJ, United States; [3]Johns Hopkins University Hospital, Baltimore, MD, United States; [4]Tulane University School of Medicine, New Orleans, LA, United States.

Introduction

- HLA molecular mismatching (HLA-MM) via HLA eplets goes beyond traditional HLA antigen mismatch for associations with transplant outcomes.

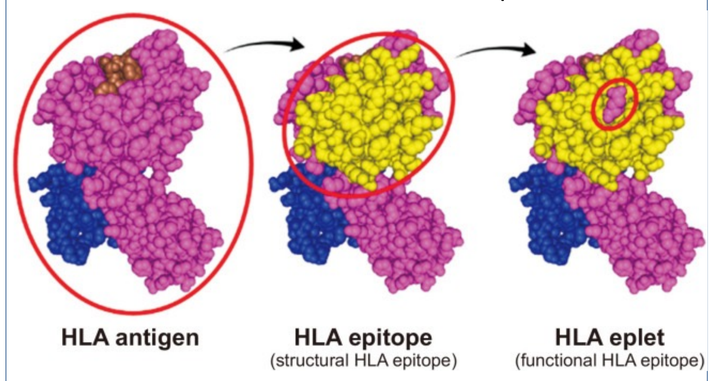


Figure 1. HLA eplets are small polymorphic amino acid motifs that may influence antibody binding.
Lee H, Oh EJ. Eplet mismatch analysis in kidney transplantation: from concept to clinical practice. *Clin Transplant Res*. 2025;39(3):183-199.

- Through the American Society for Transplantation (AST), a group of histocompatibility and clinical organ transplant experts is conducting a manual systematic review using the DistillerSR software to assess the efficacy of various HLA-MM algorithms in predicting post-kidney transplant outcomes.

- As an exploratory supplement, we evaluated Elicit, an AI service for automating research and systematic reviews.

Methods

- In the manual review, 12 reviewers (in 6 pairs; HLA typing expert + organ transplant clinician) used DistillerSR to evaluate an initial search of papers, then excluded candidate papers that did not meet criteria at abstract and full text review stages.

- We prompted Elicit AI tool in December 2025 to conduct a parallel systematic review using the same criteria (Figure 2). Elicit uses multiple large language models (LLMs) from OpenAI and Anthropic, and retrieval-augmented generation (RAG) for querying its paper database.

- Elicit AI provided abstract and full-text screening scores based on 8 specified criteria:
  - Study population;
  - HLA-MM assessment methods;
  - Comparator method (presence of antigen- or allele-level HLA-MM comparison);
  - Clinical outcomes;
  - Follow-up duration;
  - Organ type focus;
  - Study design and quality; and
  - Post-transplant outcome data.

- Elicit included papers that met a permissive default inclusion threshold of 2.5 out of 5.

Comparing the risk score of three HLA-MM assessment methods (HLA-Matchmaker, Epitope Mismatch Algorithm (EMMA), PIRCHE) against antigen (allele) level HLA mismatch assessment among pediatric and adult kidney transplant recipients, which method is more effective in predicting:

1. Post-transplant risk of T-cell mediated cellular rejection (TCMR)?
2. Post-transplant risk of antibody mediated cellular rejection (ABMR)?
3. Post-transplant kidney function measured using estimated glomerular filtration rate (eGFR)?
4. Post-transplant risk of transplant failure (graft loss)?
5. Of the three HLA-MM assessment methods (HLA-Matchmaker, Epitope Mismatch Algorithm (EMMA), PIRCHE) which performs best in predicting post-transplant outcomes?

Figure 2. Questions asked to Elicit to prompt systematic review.

Results

	Candidate papers identified	Passed abstract screen	Passed full text screen	Included in systematic review
Manual via DistillerSR	787	347	116	65
Elicit AI	500*	399	89	80*

Table 1. Comparison of papers included by Elicit and manual reviews at each screening stage. *Elicit's technical limitations capped its initial screen at 500 sources and its final systematic review at 80 sources out of 89 that met the threshold. Elicit has improved since May 2026 to screen up to 40,000 sources and extract data from up to 200.

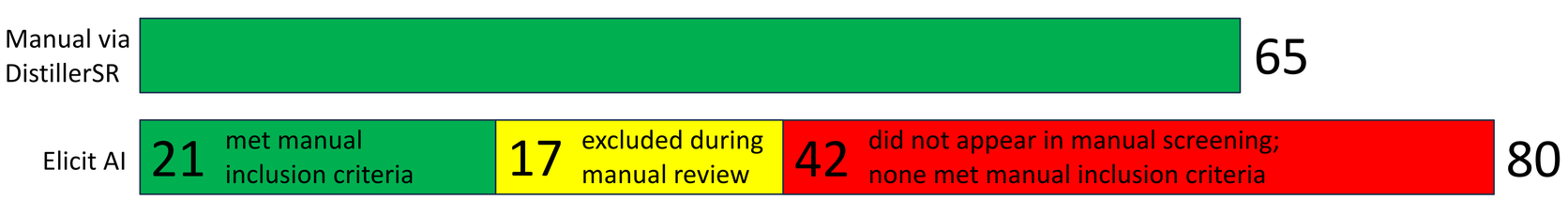


Figure 3. Comparison of papers included by Elicit and manual reviews at the final stage. Of the 65 articles that met manual screening criteria, only 21 (32.3%) were present in Elicit's 80-source systematic review. Elicit included 17 articles which were evaluated but excluded through manual review, while the remaining 42 articles did not appear in the initial manual screening. **Among these 42, Elicit misidentified many abstracts as full texts, and none of these would have met our manual inclusion criteria.**

Discussion and Conclusions

- Elicit performed poorly relative to DistillerSR with manual review for initial candidate paper selection and the subsequent screening steps.

- **Elicit missed roughly two-thirds of the studies that human experts identified as relevant, and it included many studies that did not actually meet the review's criteria.**

- The Elicit AI tool as of December 2025 was not yet reliable enough to replace trained human reviewers when selecting studies for medical research summaries, though it and other tools may still be useful for later steps such as pulling specific facts out of papers that have already been chosen.

- Because future guidance on which kidney donors are best matched to which patients depends on accumulated evidence from biomedical research studies, it is important to ensure this work is done by trained experts rather than by AI alone.

Contact

Loren Gragert
Tulane University Department of Biomedical Informatics and Genomics
lgragert@tulane.edu

Reference

Lee H, Oh EJ. Eplet mismatch analysis in kidney transplantation: from concept to clinical practice. *Clin Transplant Res*. 2025;39(3):183-199. doi:10.4285/ctr.25.0046

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com