A comparison of two artificial intelligence approaches for analyzing insights from publication citation statements

Karol Bociek,[1] Janine Dovey,[2] Hollie Rawlings,[3] Adam Errington,[2] Eileen Hartman,[4] Shailesh Desai,[4] David Lickorish,[4] Tomas Rees[3]

[1]Takeda Pharmaceuticals U.S.A., Inc., Cambridge, MA, USA; [2]PharmaGenesis Cardiff, Cardiff, UK; [3]Oxford PharmaGenesis, Oxford, UK; [4]Takeda Pharmaceuticals U.S.A., Inc., Lexington, MA, USA

Presenting author: Karol Bociek (karol.bociek@takeda.com)

Poster number 23

Scan the QR code to access the poster and any additional materials on your mobile device as well as forward yourself a link to it via email. If you do not have a QR reader, please enter https://tiny.one/mwz1959p1z in your browser.

Background

- Pharmaceutical companies and other research organizations frequently rely on quantitative publication metrics, such as citation metrics, to assess the impact of research outputs.[1]
- Citation metrics are difficult to interpret because the purpose and meaning of citations vary widely.[2]
- Citation statements (sentences containing the in-text citation from the citing article) provide context to the citation and may indicate the concepts that the citing author wishes to highlight.[3]
- The increasing accessibility of artificial intelligence (AI) tools offers an opportunity to analyze citation statements and search for themes or topics of interest, providing a deeper understanding of how publications are being received by experts in the field.

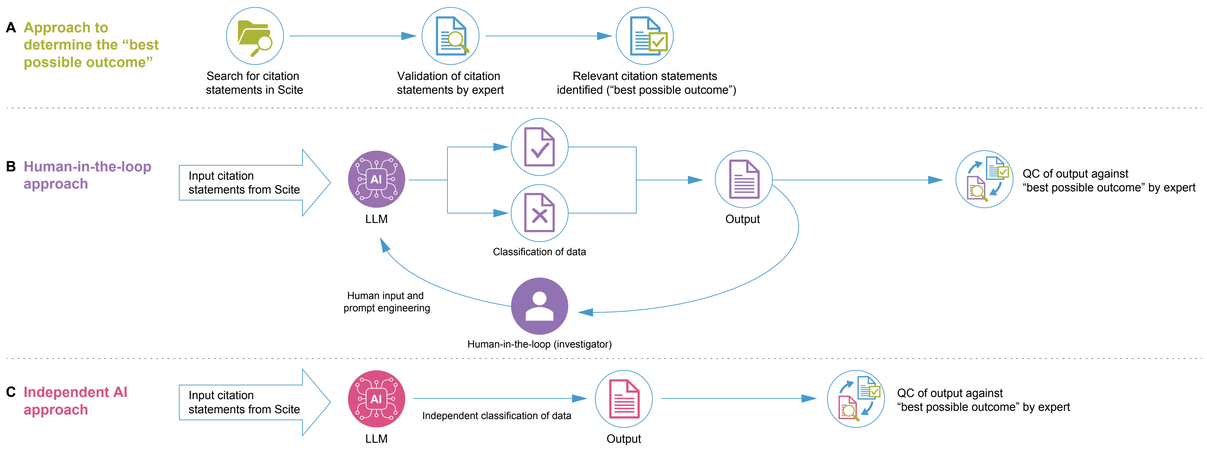
Objective

- The objective of this project was to compare the performance of two large language model (LLM) approaches in identifying relevant citation statements:
  - a human-in-the-loop approach, involving human oversight and input
  - an independent AI approach, involving a chat interface LLM.

Methods

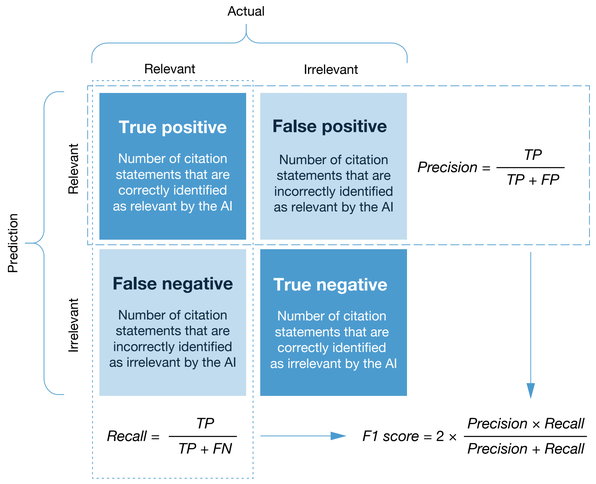
- Scite[4] was used to automatically extract citation statements for 62 Takeda-sponsored publications related to antidepressant therapy that were published between January 2022 and July 2024.
- Citation statements relevant to the predefined topic of residual symptoms of antidepressant therapy were identified by expert evaluation to determine the "best possible outcome" (**Figure 1A**).
- Using a custom AI tool that queried a secure GPT-4o endpoint, a human-in-the-loop approach was taken to analyze citation statements for their relevancy to the same predefined topic. Outputs were refined by a study investigator by prompt engineering, and the final output was quality checked by an expert against the relevant citation statements identified in the "best possible outcome" approach (**Figure 1B**).
- Citation statements were also presented via batch upload to a secure GPT-4o chat interface LLM, which was asked to identify all relevant citation statements to produce a referenced report on residual symptoms of antidepressant therapy. The report was quality checked against the validated data by an expert (**Figure 1C**).
- Confusion matrices were used to calculate precision (the measure of how often citation statements were correctly identified as relevant), recall (the measure of how many relevant citation statements were successfully retrieved), and F1 score (a balanced metric between precision and recall and an overall indicator of performance) to compare the quality of outputs from each AI approach[5] (**Figure 2**).

Figure 1. Overview of approaches to analyze citation statements.



AI, artificial intelligence; LLM, large language model; QC, quality check.

Figure 2. Performance calculations.



AI, artificial intelligence; FN, false negative; FP, false positive; TP, true positive.

Results

- Of 62 Takeda publications, 17 had citation statements, of which there were 117 in total. After the expert checked for relevancy and validated the data, 9 citation statements (7.7%) were relevant to the topic of residual symptoms of antidepressant therapy (**Figure 3**).

Figure 3. Number of relevant citation statements identified in the approach for "best possible outcome".

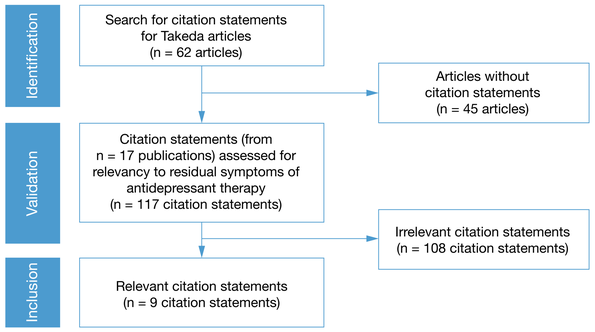
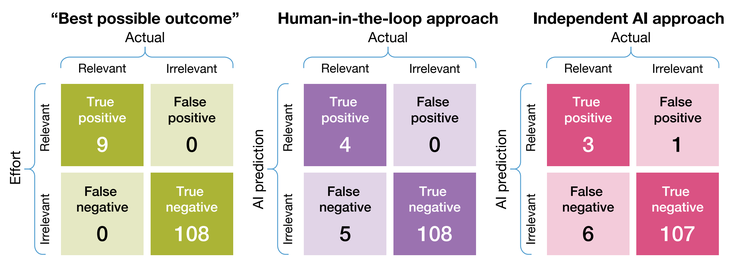


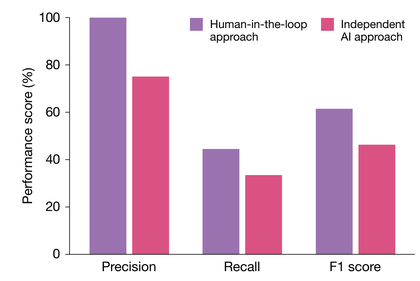
Figure 4. Confusion matrices for AI approaches.



The "best possible outcome" was determined by expert evaluation of citation statements.
AI, artificial intelligence.

- The human-in-the-loop approach correctly identified 4 (44.4%) relevant citation statements (true positives); 5 (55.6%) were missed (false negatives) and none were incorrectly identified as relevant (**Figure 4**).
- The independent AI approach correctly identified 3 (33.3%) relevant citation statements (true positives); 6 (66.6%) were missed (false negatives) and 1 citation statement was incorrectly identified as relevant (false positive) (**Figure 4**).
- Of the 4 relevant citation statements correctly identified by the human-in-the-loop approach, 3 were common with the independent AI approach. Commonly missed citation statements (false negatives) did not explicitly refer to residual symptoms and mostly referred to partial response to therapy (see supplemental materials).
- The report produced by the chat interface approach typically used other, unrelated citation statements from the dataset to extrapolate insights (see supplemental materials), whereas the human-in-the-loop approach provided focused results without extrapolating beyond the relevant information in the text.
- Performance indicators for each AI approach are shown in **Figure 5**.
  - The human-in-the-loop approach had a better precision score (100.0% vs 75.0%), recall score (44.4% vs 33.3%), and F1 score (61.5% vs 46.2%) compared with the independent AI approach.

Figure 5. Performance indicators for AI approaches.



AI, artificial intelligence.

Limitations

- Only a small number of citation statements were relevant to the predefined topic and included in the analysis.
  - While small datasets can make human validation of AI outputs more manageable, future analyses should include additional topics to allow for more statistically robust comparisons of performance between the AI approaches.
- The human investigator in this study was not a subject matter expert in antidepressant therapies. This may have limited their ability to inform and guide the decision-making by the LLM in the human-in-the-loop approach, resulting in more false negatives.

Conclusions

- Citation counts are unidimensional metrics, but citation statements can be analyzed by AI tools to provide contextual citation information.
- We found that analyzing citation statements using an LLM classifier with human input improves accuracy and mitigates extrapolations and hallucinations better than using a chat interface LLM that is presented with data by batch upload.
  - A key advantage of the human-in-the-loop approach was the ability to observe the AI's reasoning and understand its decisions. This resulted in a more transparent and iterative process that allowed human interpretation of the final output without being influenced by inferences made by AI.
- Future research should compare performance across more topics and therapy areas to further validate these findings with a larger dataset.

References

1. Myers BA, Kahn KL. Clin Transl Sci. 2021;14(5):1705-12. 2. Ioannidis JPA, et al. PLoS Biol. 2016;14(9):e1002542. 3. Nicholson JM, et al. Quant Sci Stud. 2021;2(3):882-98. 4. Brody S, et al. J Med Libr Assoc. 2021;109(4):707-10. 5. Hicks SA, et al. Sci Rep. 2022;12(1):35395867.

Disclosures

KB is an employee of Oxford PharmaGenesis Ltd, contracted to serve as Publications Lead at Takeda Pharmaceuticals U.S.A., Inc., Cambridge, MA, USA. JD and AE are employees of Oxford PharmaGenesis Ltd, Cardiff, UK. HR and TR are employees of Oxford PharmaGenesis Ltd, Oxford, UK. EH, SD, and DL are employees of Takeda Pharmaceuticals U.S.A., Inc., and hold stock and/or stock options in Takeda Pharmaceutical Company Limited.

Acknowledgments

This study was funded by Takeda Pharmaceuticals U.S.A., Inc.