# Evaluating the Effectiveness of AI-Generated vs. Real Abstracts in Training Machine Learning Models for Study Selection in Systematic Literature Reviews

Cusson E[1], Bravo À[1], Atanasov P[1]

[1]Amaris Consulting, Barcelona, Spain

## BACKGROUND

- Systematic literature reviews (SLRs) are now frequently considered the highest standard in the hierarchy of evidence. In 2019, approximately 80 SLRs were published daily by the scientific community, and this number has likely increased given the upward trend over the past two decades.[1]
- However, the process of conducting an SLR is both tedious and time-consuming, particularly during the screening of potential publications for inclusion. Each abstract must be reviewed by two independent reviewers, screening 500 abstracts is estimated to take around 8 hours.[2]
- Considering the increasing use of artificial intelligence (AI) and machine learning (ML) in scientific research, we previously tested the performace of an ML model trained with 200 human-reviewed publications to assist in the screening process. Our results demonstrated a screening performance of 92%, prompting us to explore the performance of an ML model trained entirely with AI-generated data, without any human involvement in the screening[3].

## OBJECTIVE

**This study evaluates the effectiveness of AI-generated decisions in training machine learning (ML) models for identifying of relevant publications in systematic literature reviews (SLRs).**

## METHODS

- An SLR on CAR-T therapy for multiple myeloma in Australia retrieved 989 publications from Embase and Pubmed.
- Entering PICOS criteria in the 'Custom instructions' section of ChatGPT 3.5 (free browser version), we asked the LLM to generate 50 abstracts meeting inclusion criteria and 50 with exclusion criteria. The prompt and iterations used to generate included abstracts are shown in Figure 1 bellow:

**Figure 1.** Example of prompts and iterations used to obtain the AI-generated abstracts

**Inclusion prompt**: *"Can you generate 10 titles and abstracts meeting the specified inclusion criteria using a peer-reviewed journal format, with narrative results supported by randomly generated numeral estimates? You can use examples from the web to generate different narrative structures incorporating some of the outcomes listed in the instructions."*

**Iterations**
***1#*** *"The abstracts should have the following structure: background, methods, results, conclusion"*
***2#*** *"Can you generate 10 other similar examples of title & abstracts including other criteria than the study type"*
***3#*** *"can you provide 10 other examples"* **(to get 20 abstracts)**
***4#*** *"can you provide 10 other similar examples but varying a little bit the outcomes reported according to the instructions"* **(repeat iteration to get 50 abstracts)**

- Abstracts of around 200 words with narrative results supported by randomly generated estimates followed a set structure: introduction, methods, results, and conclusion (see example in Figure 2).

**Figure 2.** Example of one AI-generated abstract

*"**Title**: Immunophenotypic Characteristics Associated with Response to Cilta-cel Therapy in Multiple Myeloma: A Retrospective Analysis*

***Introduction**: This retrospective analysis investigates immunophenotypic characteristics associated with response to cilta-cel therapy in multiple myeloma (MM) patients. **Methods**: MM patients treated with cilta-cel underwent immunophenotyping of tumor cells, and treatment responses were correlated with baseline characteristics. **Results**: Patients with high expression of B-cell maturation antigen (BCMA) on tumor cells demonstrated higher response rates to cilta-cel, with an overall response rate (ORR) of 90% compared to 60% in BCMA-low patients. **Conclusion**: Immunophenotypic profiling may help identify MM patients most likely to benefit from cilta-cel therapy, guiding personalized treatment strategies."*

## METHODS (CONTINUED…)

- We trained ML models with a set of abstracts to provide a relevance score to the remaining publications, organizing them to prioritize the most relevant for inclusion. Four scenarios were proposed:

**Scenario A: Human decisions**
- Trained with 100 real abstracts randomly selected and annotated by experts

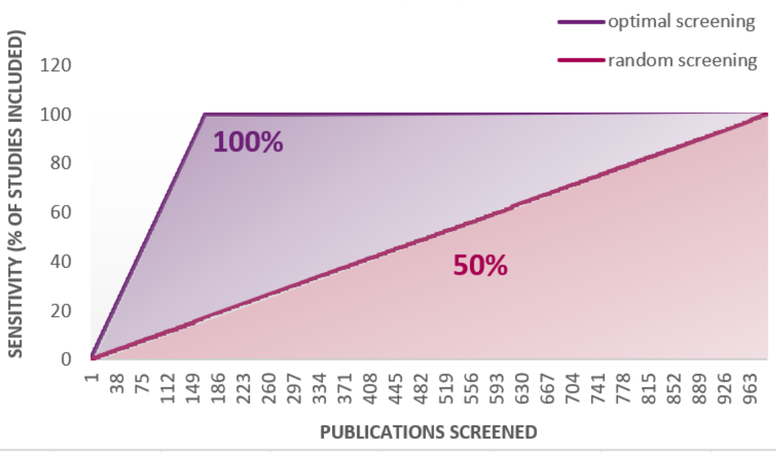
**Scenario B: AI decisions**
- Trained with 100 AI-generated abstracts (50 for inclusion and 50 for exclusion)

**Scenario C: Combined decisions**
- Trained with 100 real abstracts (scenario A) enriched with 50 AI-generated inclusion abstracts

**Scenario D: AI-derived decisions**
- Trained with top 100 real abstracts based on scores from scenario B

- Results of the logistic regression model were plotted on screening progression curves, showing the percentage of included publications found versus the percentage of publications screened, allowing us to calculate performance based on the Area Under the Curve (AUC). The curves from the four scenarios were compared with optimal screening (100%) where included publications appear first and manual screening (50%), where publications appear in random order (Figure 3).

**Figure 3.** Optimal vs Random screening progress




## RESULTS

- Scenario A achieved a performance of 79.51%. Scenario B demonstrated 74.48%. Scenario C showed the highest performance, reaching 81.49%. Scenario D achieved 79.86% (see Figures 4 to 7).
- Scenario C identified 80% of the included publications by screening only 50% of the total set, outperforming scenarios A and D, which required screening 54% and 59% of the publications, respectively. Scenario B needed to screen 69% to identify 80% of the included publications.

**Figure 4.** Scenario A screening progress

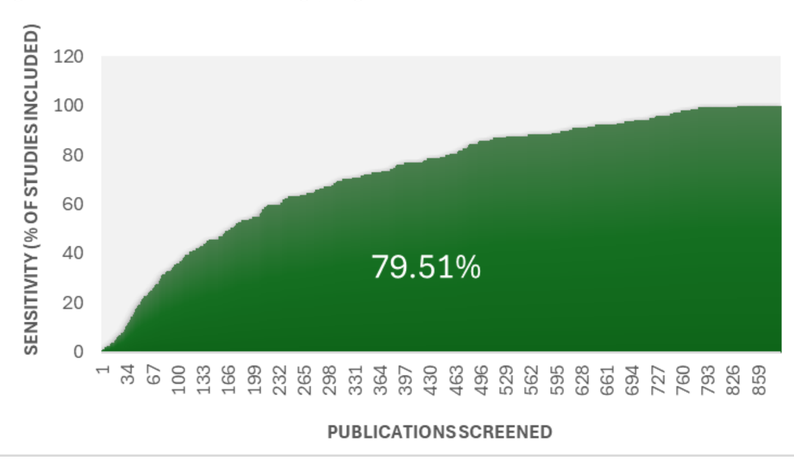
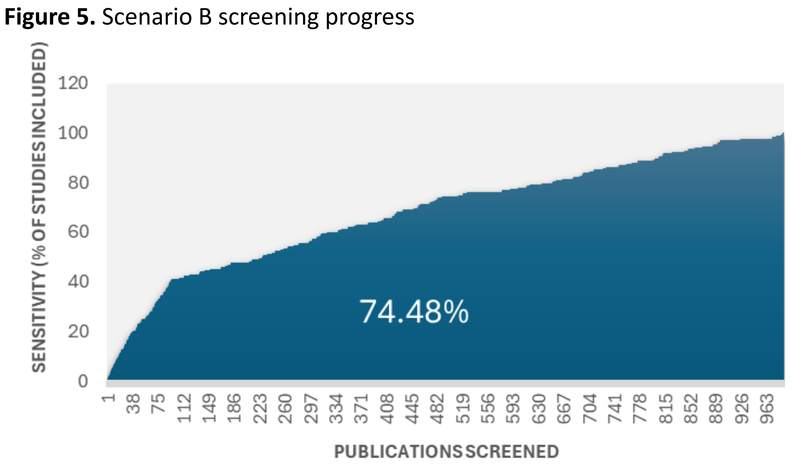



**Figure 5.** Scenario B screening progress




## RESULTS (CONTINUED…)

**Figure 6.** Scenario C screening progress

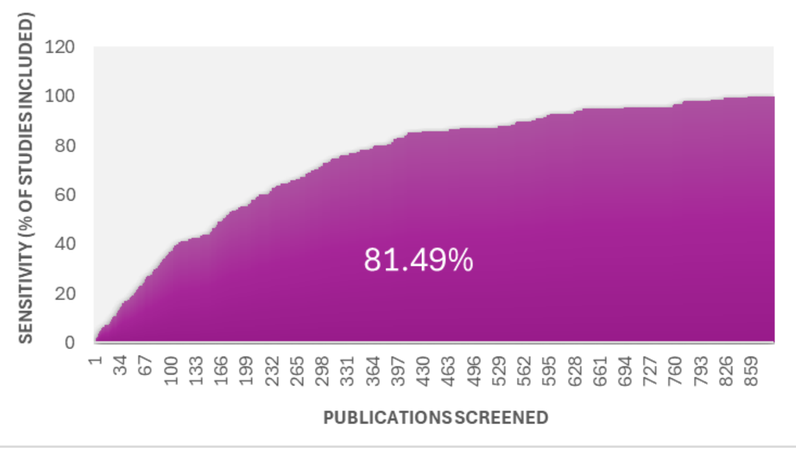
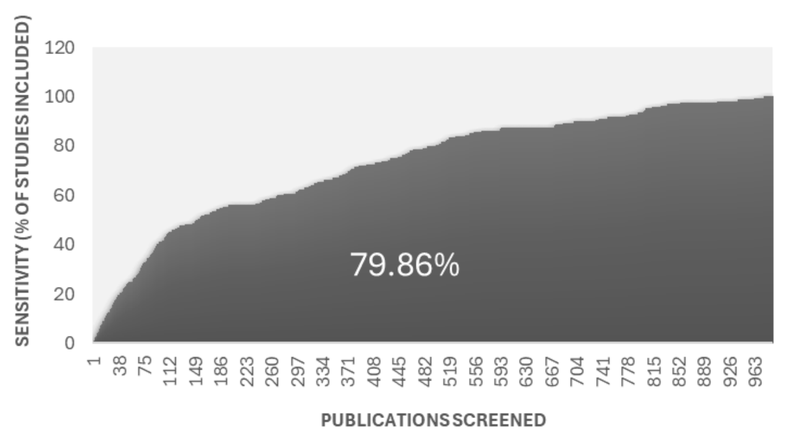



**Figure 7.** Scenario D screening progress




## DISCUSSION & CONCLUSION

- Our study used ChatGPT3.5 to generate data for training machine learning models, which allowed model training to start sooner without requiring large amounts of real data upfront.
- Combining real data with some AI-generated publications in scenario C yielded better outcomes compared to analyzing publications in random order. Adding AI-generated publications helped balance the data, resulting in a more accurate model.
- Interestingly, we found that a completely made-up dataset could be effectively used to achieve good results even before beginning the screening process (scenario D). This represents a novel application of AI for quickly reviewing a large volume of studies.
- By using artificial data for early model training, significant time and effort can be saved during the review process. These findings suggest machine learning could serve as a second reviewer in systematic literature reviews, improving efficiency while preserving accuracy and rigor by keeping one human reviewer involved.

### *Limitations*

- Generating abstracts with ChatGPT remains time consuming and needs the human in the loop to verify the quality of the abstracts.
- 4 iterations were needed along the process to generate the 50 abstracts with ChatGPT. Additionally, ChatGPT tends to get tired, and we had to generate the abstracts per batches of 10. When requesting a higher number of abstracts (e.g. 25), ChatGPT started to provide much shorter and similar samples. We also had to remind it to vary the numerical estimates and the narrative within abstracts.

## REFERENCES

1. Hoffmann F, Allers K, Rombey T, Helbach J, Hoffmann A, Mathes T, Pieper D. Nearly 80 systematic reviews were published each day: Observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol. 2021 Oct;138:1-11. doi: 10.1016/j.jclinepi.2021.05.022. Epub 2021 Jun 4. PMID: 34091022.
2. Waffenschmidt, S., Knelangen, M., Sieben, W. *et al.* Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med Res Methodol* **19**, 132 (2019). https://doi.org/10.1186/s12874-019-0782-0
3. Bravo, À., Patel, P., & Atanasov, P. (2023). MSR63 Implementing Simple Active Learning (AL) Boosters Considerably Improves the Early Identification of Relevant Studies in the Systematic Literature Review (SLR) Process. Value in Health, 26(12), S404-S405.