Is GPT-4o capable of automating detailed data extraction for Systematic Literature Reviews (SLRs)?

Bravo À[1], Cusson E[1], Shalaby N.[2], Atanasov P[1]

[1]Amaris Consulting, Barcelona, Spain; [2]Amaris Consulting, Toronto, Canada

INTRODUCTION

- Systematic Literature Reviews (SLRs) synthesize evidence to guide clinical decisions and health policy. **Data extraction, a detailed and labor-intensive task**, is often a bottleneck in the review process [1].
- Recent AI models, like GPT-4o, offer new opportunities **to automate data extraction** by processing complex information with minimal human input [2]. GPT-4o demonstrates potential in zero-shot learning, handling tasks without prior specific training.
- **This study evaluates GPT-4o's ability** to extract detailed data from NSCLC trial publications, potentially streamlining SLRs by saving time and resources while achieving high accuracy.

OBJECTIVES

- **Accuracy assessment**: Measure GPT-4o's precision in extracting study design elements from clinical trials.
- **Strengths and limitations:** Identify areas where GPT-4o performs well or encounters challenges due to data complexity.
- **Efficiency gains:** Compare time savings achieved with GPT-4o versus manual extraction.
- **Future applications:** Explore how AI-driven extraction can streamline SLRs and support scalable data processing in clinical research.

METHODS

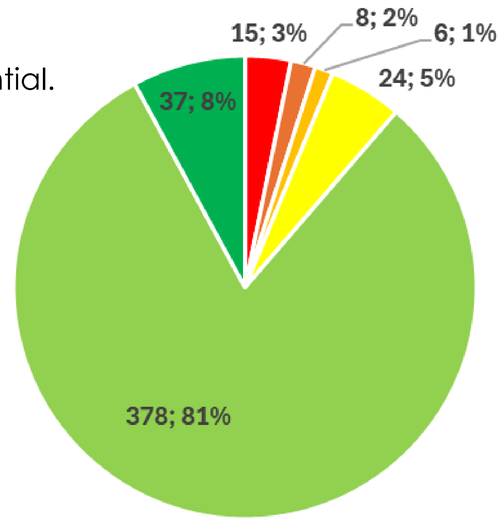
- **Data Source Selection:** We used 12 clinical trial publications on NSCLC interventions, with quality-checked information to benchmark GPT-4o's performance.
- **Data entry and template definition:** Defined 39 study design data elements, resulting in 468 data points, organized into a standardized template for consistent extraction.
- **Text conversion and prompt preparation:** Publications were converted from PDF to text using Python, then processed with prompts containing extraction instructions to guide GPT-4o.
- **API integration and data processing:** The prepared text and prompts were sent to GPT-4o via an Azure-hosted API, with extracted data automatically structured for easy comparison.
- **Accuracy assessment:** GPT-4o's extracted data were compared to manual entries, with discrepancies classified as significant, minor, fabricated, or partially missing data.
- **Error analysis and contextual complexity:** Errors were analyzed, focusing on complex entries (e.g., subgroup analyses), highlighting GPT-4o's interpretive challenges.
- **Efficiency measurement:** Finally, we recorded the time taken by GPT-4o to extract data elements per publication and calculated the average extraction time. This metric was then compared to manual extraction times to evaluate the model's efficiency gains.

Group	Data Element	Human Response	GPT4o Response
Trial Characteristics	Study start - Completion date	9/19/2014 - May 9, 2016	September 19, 2014 to October 29, 2015
	Countries Included Specify China only studies	Australia, Austria, Belgium, Canada, Denmark , Finland, France, Germany, Ireland, Israel, Italy, Japan, Netherlands, Spain, United Kingdom, United States,	'United States', 'Spain', 'Canada', 'Australia', 'Austria', 'Germany', 'Israel', 'Italy', 'Japan'
Follow-up and Data cut-off	Length of follow-up	6 weeks	Median follow-up time was 28.1 months for B+CP and 26.9 months for PI+CP.
	Data cut-off date	Not Reported	November 8, 2017
Other details	Frequency of outcome assessment	Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
		Not Reported	Every 6 weeks from day 1 of cycle 1 for the first 48 weeks, and every 9 weeks thereafter until the occurrence of disease progression.
		Tumour imaging by CT (preferred) or MRI was done at baseline, every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.	Every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.
		Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
		Not Reported	Radiographic tumour imaging was done at baseline and scheduled for every 9 weeks for the first 45 weeks, then every 12 weeks thereafter.
		Not Reported	Tumor imaging was scheduled for weeks 6 and 12, then every 9 weeks through week 48 and every 12 weeks thereafter.
		Not Reported	Every 6 weeks (± 7 days) for the first 48 weeks following cycle 1, day 1; after 48 weeks, every 9 weeks (± 7 days).
	Age	≥18 years	62.5 years (median, pembrolizumab plus chemotherapy group); 63.2 years (median, chemotherapy group)
	Exclusion criteria	Patients were ineligible if they were receiving systemic glucocorticoids (excluding daily glucocorticoid- replacement therapy for conditions such as adrenal or pituitary insufficiency) or other immunosuppressive treatment or if they had untreated brain metastases, active autoimmune disease for which they had received systemic treatment during the previous 2 years, active interstitial lung disease, or a history of pneumonitis for which they had received glucocorticoids.	Patients receiving systemic glucocorticoids or other immunosuppressive treatment, untreated brain metastases, active autoimmune disease, active interstitial lung disease, or a history of pneumonitis.
	Subsequent therapies allowed	YES	No crossover to the atezolizumab group was permitted.

Table 1. ***Examples of data extraction by human reviewers and GPT-4o, showing superior (green) and correct responses (light green), minor discrepancies (yellow), fabricated response (light orange), moderate (orange) and significant errors (red).***

RESULTS

- GPT-4o achieved an **88.7% accuracy** rate, successfully extracting 415 out of 468 data elements (Table 1 shows some examples).
- In **37 instances**, GPT-4o provided **more detailed information** than the manual extraction process, illustrating its zero-shot learning potential.
- In **53 cases**, elements were generated **erroneously**:
  - **Significant Errors (15):** Misinterpretations impacting data integrity, where GPT-4o misunderstood specific data elements or contexts.
  - **Minor Errors (8):** Small formatting or wording inconsistencies that did not affect overall accuracy.
  - **Fabricated Data (6):** Instances where GPT-4o generated data not present in the source, indicating over-generalization.
  - **Partially Missing Data (24):** Missing information in responses, often linked to entries requiring complex contextual understanding.
- A **significant concentration of errors** (29) was observed in 8 data elements that require complex contextual understanding, such as subgroup analyses and details of therapeutic protocols.
- The pipeline averaged **27.75 seconds per publication**, a notable time savings versus manual extraction.



CONCLUSIONS

- **GPT-4o demonstrates strong potential in automating data extraction for SLRs**, achieving an 88.7% accuracy rate and significantly reducing manual workload. Its zero-shot learning capabilities allow it to handle structured data effectively, making it valuable for large-scale reviews where efficiency is key.
- **GPT-4o struggles with contextually complex data elements**, such as subgroup analyses and detailed protocols, where errors were more frequent. This suggests a need for further refinement, especially for data requiring specialized understanding.
- **GPT-4o's automation capabilities can accelerate** evidence synthesis for clinical and policy applications. With additional fine-tuning, the model could perform even better on complex study designs, expanding its utility in SLRs.

REFERENCES

1. Mahmoudi, H., Chang, D., Lee, H., Ghaffarzadegan, N., & Jalali, M. S. (2024). A Critical Assessment of Large Language Models for Systematic Reviews: Utilizing ChatGPT for Complex Data Extraction. Available at SSRN 4797024.
2. Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods.

2

Backup

Is GPT-4o capable of automating detailed data extraction for Systematic Literature Reviews (SLRs)?

Bravo À[1], Cusson E[1], Shalaby N.[2], Atanasov P[1]

[1]Amaris Consulting, Barcelona, Spain; [2]Amaris Consulting, Toronto, Canada

INTRODUCTION

- Systematic Literature Reviews (SLRs) synthesize evidence to guide clinical decisions and health policy. Data extraction, a detailed and labor-intensive task, is often a bottleneck in the review process.
- Recent AI models, like GPT-4o, offer new opportunities to automate data extraction by processing complex information with minimal human input. GPT-4o demonstrates potential in zero-shot learning, handling tasks without prior specific training.
- This study evaluates GPT-4o's ability to extract detailed data from NSCLC trial publications, potentially streamlining SLRs by saving time and resources while achieving high accuracy.

OBJECTIVES

- **Accuracy assessment**: Measure GPT-4o's precision in extracting study design elements from clinical trials.
- **Strengths and limitations:** Identify areas where GPT-4o performs well or encounters challenges due to data complexity.
- **Efficiency gains:** Compare time savings achieved with GPT-4o versus manual extraction.
- **Future applications:** Explore how AI-driven extraction can streamline SLRs and support scalable data processing in clinical research.

METHODS

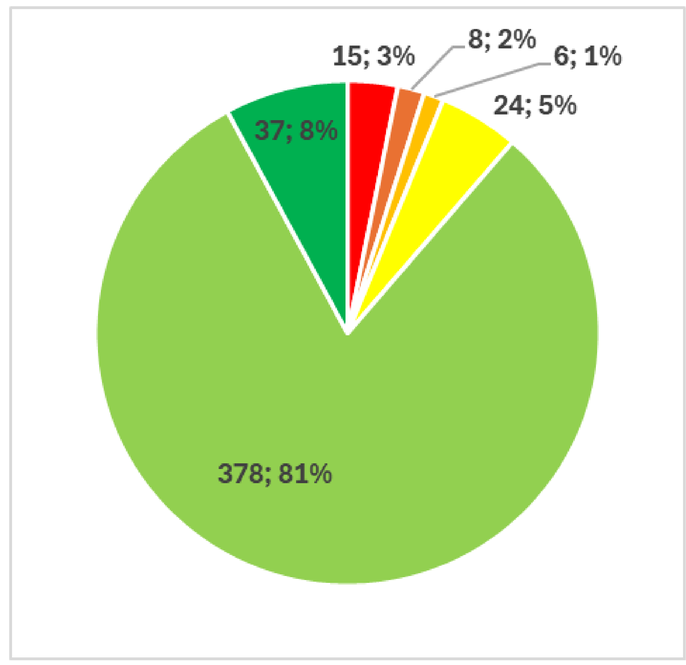
- **Data Source Selection:** We used 12 clinical trial publications on NSCLC interventions, with quality-checked information to benchmark GPT-4o's performance.
- **Data entry and template definition:** Defined 39 study design data entries, resulting in 468 data points, organized into a standardized template for consistent extraction.
- **Text conversion and prompt preparation:** Publications were converted from PDF to text using Python, then processed with prompts containing extraction instructions to guide GPT-4o.
- **API integration and data processing:** The prepared text and prompts were sent to GPT-4o via an Azure-hosted API, with extracted data automatically structured for easy comparison.
- **Accuracy assessment:** GPT-4o's extracted data were compared to manual entries, with discrepancies classified as significant, minor, fabricated, or partially missing data.
- **Error analysis and contextual complexity:** Errors were analyzed, focusing on complex entries (e.g., subgroup analyses), highlighting GPT-4o's interpretive challenges
- **Efficiency measurement:** Finally, we recorded the time taken by GPT-4o to extract data elements per publication and calculated the average extraction time. This metric was then compared to manual extraction times to evaluate the model's efficiency gains.

Group	Data Element	Human Response	GPT4o Respone
Trial Characteristics	Study start - Completion date	9/19/2014 - May 9, 2016	September 19, 2014 to October 29, 2015
	Countries Included Specify China only studies	Australia, Austria, Belgium, Canada, Denmark , Finland, France, Germany, Ireland, Israel, Italy, Japan, Netherlands, Spain, United Kingdom, United States,	'United States', 'Spain', 'Canada', 'Australia', 'Austria', 'Germany', 'Israel', 'Italy', 'Japan'
Follow-up and Data cut-off	Length of follow-up	6 weeks	Median follow-up time was 28.1 months for B+CP and 26.9 months for PI+CP.
	Data cut-off date	Not Reported	November 8, 2017
Other details	Frequency of outcome assessment	Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
		Not Reported	Every 6 weeks from day 1 of cycle 1 for the first 48 weeks, and every 9 weeks thereafter until the occurrence of disease progression.
		Tumour imaging by CT (preferred) or MRI was done at baseline, every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.	Every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.
		Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
		Not Reported	Radiographic tumour imaging was done at baseline and scheduled for every 9 weeks for the first 45 weeks, then every 12 weeks thereafter.
		Not Reported	Tumor imaging was scheduled for weeks 6 and 12, then every 9 weeks through week 48 and every 12 weeks thereafter.
		Not Reported	Every 6 weeks (± 7 days) for the first 48 weeks following cycle 1, day 1; after 48 weeks, every 9 weeks (± 7 days).
	Age	≥18 years	62.5 years (median, pembrolizumab plus chemotherapy group); 63.2 years (median, chemotherapy group)
	Exclusion criteria	Patients were ineligible if they were receiving systemic glucocorticoids (excluding daily glucocorticoid- replacement therapy for conditions such as adrenal or pituitary insufficiency) or other immunosuppressive treatment or if they had untreated brain metastases, active autoimmune disease for which they had received systemic treatment during the previous 2 years, active interstitial lung disease, or a history of pneumonitis for which they had received glucocorticoids.	Patients receiving systemic glucocorticoids or other immunosuppressive treatment, untreated brain metastases, active autoimmune disease, active interstitial lung disease, or a history of pneumonitis.
	Subsequent therapies allowed	YES	No crossover to the atezolizumab group was permitted.

RESULTS

GPT-4o displayed promising accuracy and efficiency in automating data extraction for SLRs, with the following key findings:

- **Overall Accuracy:**
  - Achieved an 88.7% accuracy rate, successfully extracting 415 out of 468 data elements.
  - Demonstrated strong capability in accurately processing structured data with minimal human intervention.
  - In 37 instances, GPT-4o provided more detailed information than the manual extraction process, illustrating its zero-shot learning potential.
  - Added value by inferring details that were sometimes overlooked in manual extraction, potentially enhancing data quality in SLRs.
- **Error Analysis:** A total of 53 errors were observed, categorized as follows:
  - **Significant Errors (15):** Misinterpretations impacting data integrity, where GPT-4o misunderstood specific data elements or contexts.
  - **Minor Errors (8):** Small formatting or wording inconsistencies that did not affect overall accuracy.
  - **Fabricated Data (6):** Instances where GPT-4o generated data not present in the source, indicating over-generalization.
  - **Partially Missing Data (24):** Missing information in responses, often linked to entries requiring complex contextual understanding.
- **Contextual Complexity and Error Distribution:** Errors clustered in fields needing nuanced understanding (e.g., subgroup analyses), indicating limits in handling context-heavy data.
- **Efficiency:** Averaged 27.75 seconds per publication, a notable time savings versus manual extraction.
- **SLR Automation Potential:** GPT-4o shows significant promise for streamlining SLRs, with efficiency gains and reduced workload, though complex data may require further model refinement



Is GPT-4o capable of automating detailed data extraction for Systematic Literature Reviews (SLRs)?

Bravo À[1], Cusson E[1], Shalaby N.[2], Atanasov P[1]

[1]Amaris Consulting, Barcelona, Spain; [2]Amaris Consulting, Toronto, Canada

BACKGROUND

- Systematic Literature Reviews (SLRs) play a critical role in synthesizing evidence across studies, guiding clinical decisions, and shaping health policy. At the heart of this process is data extraction—the detailed task of identifying and pulling relevant information from each study, such as study design elements, participant characteristics, interventions, and outcomes. This process is inherently labor-intensive, requiring expert time, attention to detail, and substantial resources, often making it a bottleneck in the systematic review process.Recent advances in generative artificial intelligence (AI) models, like OpenAI's GPT-4, have opened new pathways for automating such labor-intensive tasks. These models can process complex textual information, potentially extracting structured data with minimal human intervention. GPT-4o, a model tuned for task-specific outputs, has shown promise in zero-shot learning, where it can understand and respond to new data requests without explicit prior training on the specific task.This study aims to assess the capabilities of GPT-4o in performing detailed data extraction from clinical trial publications. By focusing on metastatic non-small cell lung cancer (NSCLC) studies, we examine whether GPT-4o can elements, often requiring nuanced contextual understanding. The potential for using GPT-4o in SLRs is significant, as it could streamline data extraction processes, saving researchers time and resources while maintaining high levels of accuracy.

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

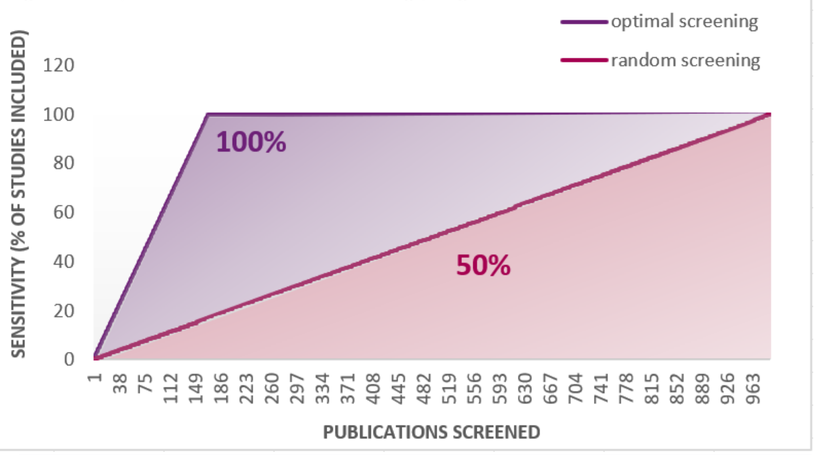
Scenario B: AI decisions
- Trained with 100 AI-generated abstracts (50 for inclusion and 50 for exclusion)

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

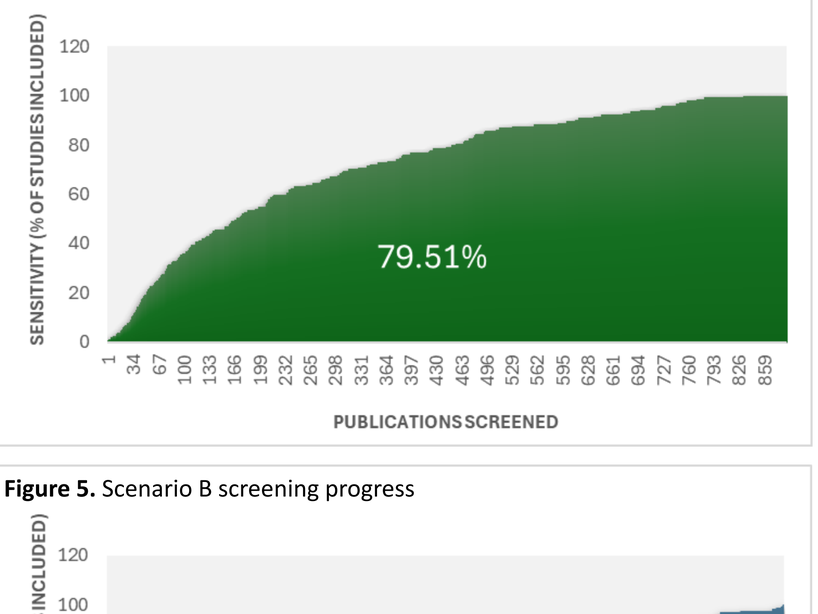
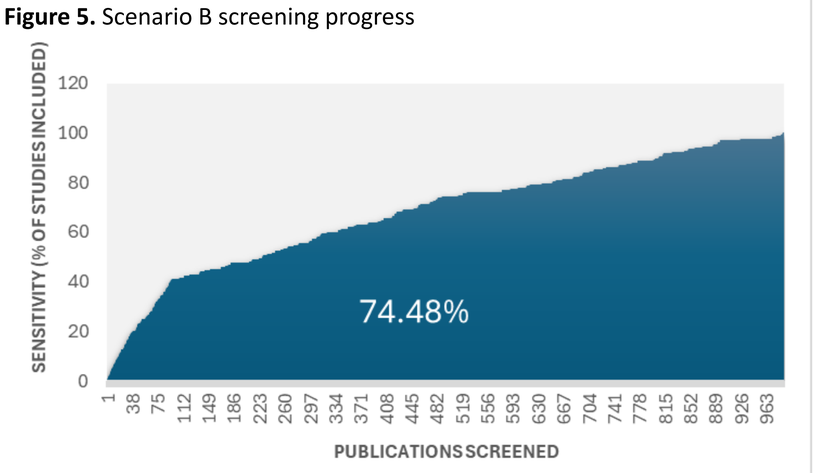


Figure 5. Scenario B screening progress



RESULTS (CONTINUED…)

Figure 6. Scenario C screening progress

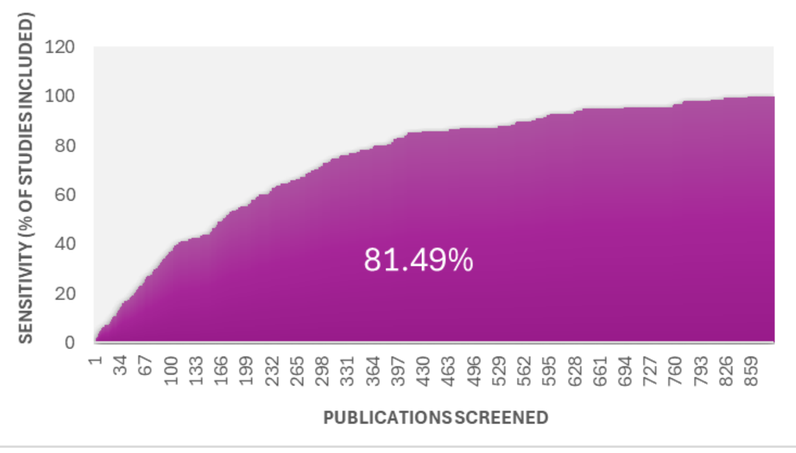
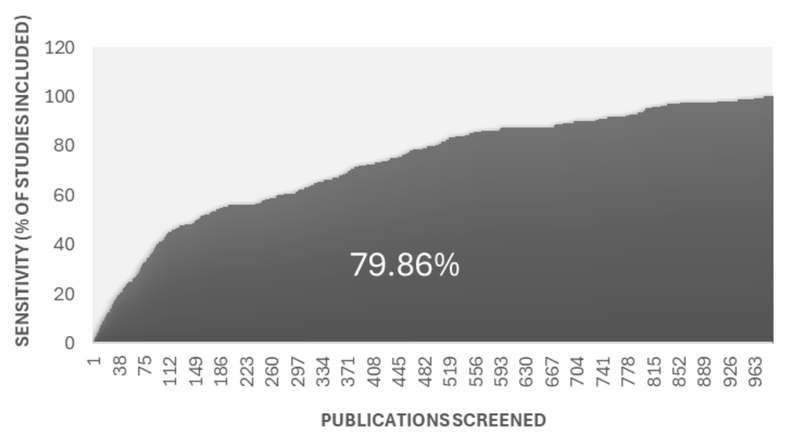


Figure 7. Scenario D screening progress



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

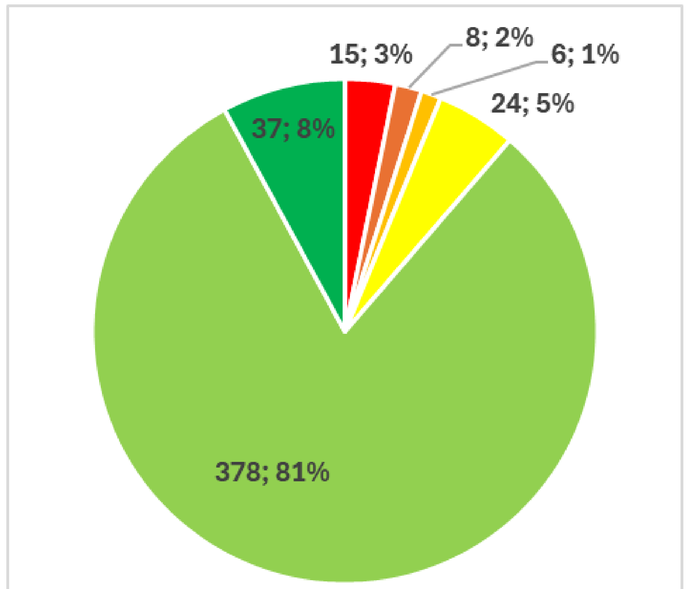
- Generating abstracts with ChatGPT remains time consuming and needs the human in the loop to verify the quality of the abstracts.

- 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

Error Analysis:
- Study Design:
  - 39 Data Types
  - 12 Trials
  - 468 Data Elements
    - 415 correct responses
    - 53 erroneous responses



Types of errors	Description
Significant Error	This error severely affects the accuracy of the data and, if not corrected, could result in incorrect conclusions; examples include major miscalculations or incorrect data assignments.
Moderate Error	This error is less severe than a significant error but still degrades the quality of the data; examples include minor miscalculations or rounding errors that do not greatly impact the overall usefulness of the data.
Fabricated Data	Data that appears to be invented by the language model rather than based on actual information.
Missing or Overlooked Data	Data that were present in the original source (or reference document) but were either missed or not included by the language model.
Accurate	The generated output is consistent with the human-provided response.
Superior	The generated output is better than the human-provided response.

Group	Element Type	Human Response	GPT4o Respone
Trial Characteristics	Study start - Completion date	9/19/2014 - May 9, 2016	September 19, 2014 to October 29, 2015
	Countries Included Specify China only studies	Australia, Austria, Belgium, Canada, Denmark , Finland, France, Germany, Ireland, Israel, Italy, Japan, Netherlands, Spain, United Kingdom, United States,	['United States', 'Spain', 'Canada', 'Australia', 'Austria', 'Germany', 'Israel', 'Italy', 'Japan']
Follow-up and Data cut-off	Length of follow-up	6 weeks	Median follow-up time was 28.1 months for B+CP and 26.9 months for PI+CP.
	Data cut-off date	Not Reported	November 8, 2017
Other details	Frequency of outcome assessment	Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
	Frequency of outcome assessment	Not Reported	Every 6 weeks from day 1 of cycle 1 for the first 48 weeks, and every 9 weeks thereafter until the occurrence of disease progression.
	Frequency of outcome assessment	Tumour imaging by CT (preferred) or MRI was done at baseline, every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.	Every 6 weeks for the first 18 weeks, then every 9 weeks through the first 12 months and every 12 weeks thereafter.
	Frequency of outcome assessment	Not Reported	After random assignment, assessments were performed at the end of every second treatment cycle until first progression. One final tumor scan was performed 4 to 6 weeks after the first progression.
	Frequency of outcome assessment	Not Reported	Radiographic tumour imaging was done at baseline and scheduled for every 9 weeks for the first 45 weeks, then every 12 weeks thereafter.
	Frequency of outcome assessment	Not Reported	Tumor imaging was scheduled for weeks 6 and 12, then every 9 weeks through week 48 and every 12 weeks thereafter.
			Every 6 weeks (± 7 days) for the first 48 weeks following cycle 1, day 1; after 48 weeks, every 9 weeks (± 7 days).
	Age	≥18 years	62.5 years (median, pembrolizumab plus chemotherapy group); 63.2 years (median, chemotherapy group)
	Exclusion criteria	Patients were ineligible if they were receiving systemic glucocorticoids (excluding daily glucocorticoid-replacement therapy for conditions such as adrenal or pituitary insufficiency) or other immunosuppressive treatment or if they had untreated brain metastases, active autoimmune disease for which they had received systemic treatment during the previous 2 years, active interstitial lung disease, or a history of pneumonitis for which they had received glucocorticoids.	Patients receiving systemic glucocorticoids or other immunosuppressive treatment, untreated brain metastases, active autoimmune disease, active interstitial lung disease, or a history of pneumonitis.
	Subsequent therapies allowed	YES	No crossover to the atezolizumab group was permitted.

Is GPT-4o capable of automating detailed data extraction for Systematic Literature Reviews (SLRs)?

Bravo À[1], Cusson E[1], Shalaby N.[2], Atanasov P[1]

[1]Amaris Consulting, Barcelona, Spain; [2]Amaris Consulting, Toronto, Canada

INTRODUCTION

- Systematic Literature Reviews (SLRs) play a critical role in synthesizing evidence across studies, guiding clinical decisions, and shaping health policy. At the heart of this process is data extraction—the detailed task of identifying and pulling relevant information from each study, such as study design elements, participant characteristics, interventions, and outcomes. This process is inherently labor-intensive, requiring expert time, attention to detail, and substantial resources, often making it a bottleneck in the systematic review process.
- Recent advances in generative artificial intelligence (AI) models, like OpenAI's GPT-4, have opened new pathways for automating such labor-intensive tasks. These models can process complex textual information, potentially extracting structured data with minimal human intervention. GPT-4o, a model tuned for task-specific outputs, has shown promise in zero-shot learning, where it can understand and respond to new data requests without explicit prior training on the specific task.
- This study aims to assess the capabilities of GPT-4o in performing detailed data extraction from clinical trial publications. By focusing on metastatic non-small cell lung cancer (NSCLC) studies, we examine whether GPT-4o can accurately identify and retrieve complex study design elements, often requiring nuanced contextual understanding. The potential for using GPT-4o in SLRs is significant, as it could streamline data extraction processes, saving researchers time and resources while maintaining high levels of accuracy.

OBJECTIVES

This study aims to assess GPT-4o's effectiveness in automating data extraction for SLRs. Key objectives include:

- Accuracy Assessment: Measure GPT-4o's precision in extracting study design elements from clinical trials.
- Strengths and Limitations: Identify areas where GPT-4o performs well or encounters challenges due to data complexity.
- Efficiency Gains: Compare time savings achieved with GPT-4o versus manual extraction.
- Future Applications: Explore how AI-driven extraction can streamline SLRs and support scalable data processing in clinical research.

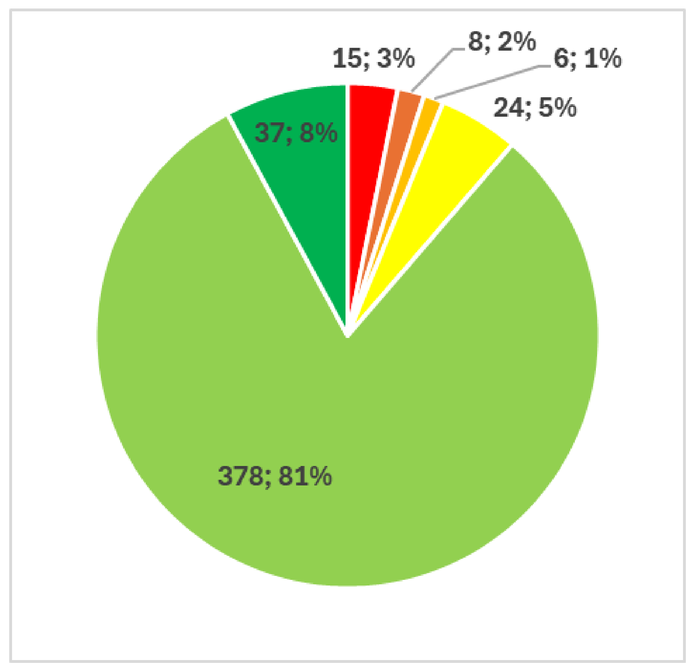
METHODS

- Data Source Selection: We used a sample of 12 clinical trial publications focused on metastatic non-small cell lung cancer (NSCLC) interventions. Each publication had previously extracted and quality-checked information to serve as a baseline for assessing GPT-4o's performance.
- Data Entry and Template Definition: We identified 39 data entries across various study design elements, including participant criteria, intervention specifics, and outcomes, resulting in a total of 468 individual data points. These data points were organized in a standardized extraction template for consistency.
- Text Conversion and Prompt Preparation: Using Python scripts, we converted the publications from PDF to text format. This allowed us to process the content efficiently and integrate it with predefined prompts. Each prompt contained detailed instructions specifying the type and format of data entries to be extracted, providing GPT-4o with clear expectations for the task.
- API Integration and Data Processing: The processed text, along with the data extraction prompt, was submitted to GPT-4o via an API hosted on Azure. The model's output was automatically structured into the data extraction template to ensure consistency and ease of comparison with the manual extraction.
- Accuracy Assessment: Extracted data from GPT-4o were compared with manually extracted data entries. Discrepancies were categorized into four types—significant errors, minor errors, fabricated data, and partially missing data—to analyze areas of strength and limitations.
- Error Analysis and Contextual Complexity: We conducted a focused analysis of GPT-4o's errors, particularly on data entries requiring complex contextual understanding (e.g., subgroup analyses, therapeutic protocols). This analysis provided insights into the types of information GPT-4o finds challenging to interpret accurately.
- Efficiency Measurement: Finally, we recorded the time taken by GPT-4o to extract data elements per publication and calculated the average extraction time. This metric was then compared to manual extraction times to evaluate the model's efficiency gains.

RESULTS

GPT-4o displayed promising accuracy and efficiency in automating data extraction for SLRs, with the following key findings:

- **Overall Accuracy:**
  - Achieved an 88.7% accuracy rate, successfully extracting 415 out of 468 data elements.
  - Demonstrated strong capability in accurately processing structured data with minimal human intervention.
  - In 37 instances, GPT-4o provided more detailed information than the manual extraction process, illustrating its zero-shot learning potential.
  - Added value by inferring details that were sometimes overlooked in manual extraction, potentially enhancing data quality in SLRs.
- **Error Analysis:** A total of 53 errors were observed, categorized as follows:
  - **Significant Errors (15):** Misinterpretations impacting data integrity, where GPT-4o misunderstood specific data elements or contexts.
  - **Minor Errors (8):** Small formatting or wording inconsistencies that did not affect overall accuracy.
  - **Fabricated Data (6):** Instances where GPT-4o generated data not present in the source, indicating over-generalization.
  - **Partially Missing Data (24):** Missing information in responses, often linked to entries requiring complex contextual understanding.
- **Contextual Complexity and Error Distribution:** Errors clustered in fields needing nuanced understanding (e.g., subgroup analyses), indicating limits in handling context-heavy data.
- **Efficiency:** Averaged 27.75 seconds per publication, a notable time savings versus manual extraction.
- **SLR Automation Potential:** GPT-4o shows significant promise for streamlining SLRs, with efficiency gains and reduced workload, though complex data may require further model refinement



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

INTRODUCTION

- 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 accuracy of an ML model trained with 100 human-reviewed publications to assist in the screening process. Our results demonstrated a 95% accuracy rate, prompting us to explore the performance of an ML model trained entirely with AI-generated data, without any human involvement in the screening (ref ABR).

OBJECTIVES

- **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."*

- 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

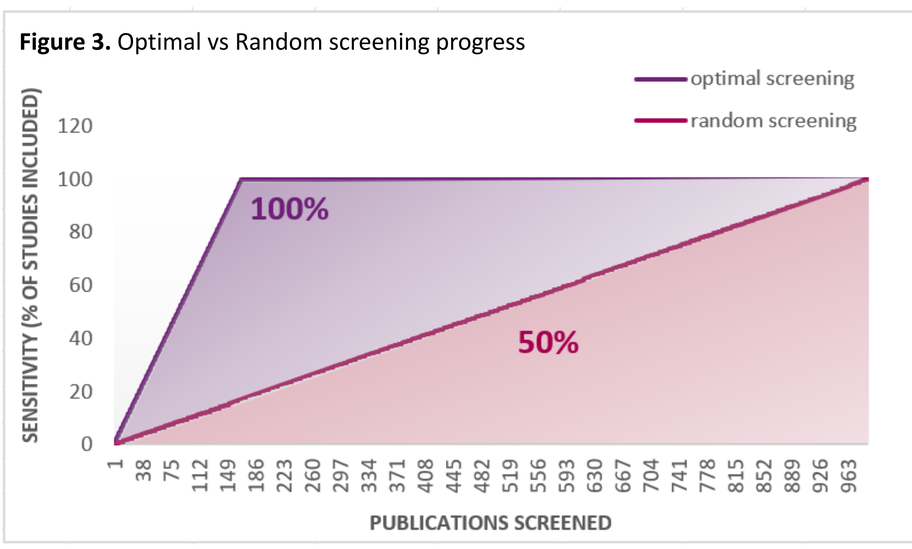
DISCLOSURES

None.

METHODS (CONTINUED…)

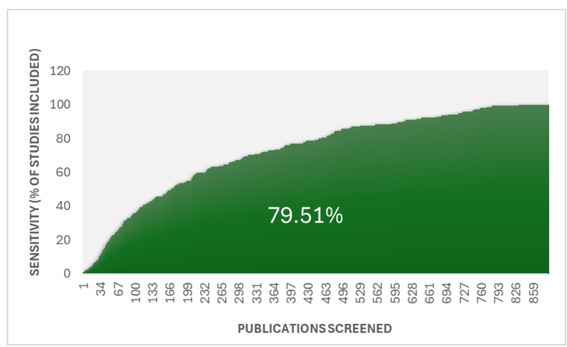
- Results of the logistic regression model were plotted on screening progression curves, showing the percentage of included publications found versus the percentage of publications screened, allowing us to calculate performance based on the Area Under the Curve (AUC). The curves from the four scenarios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

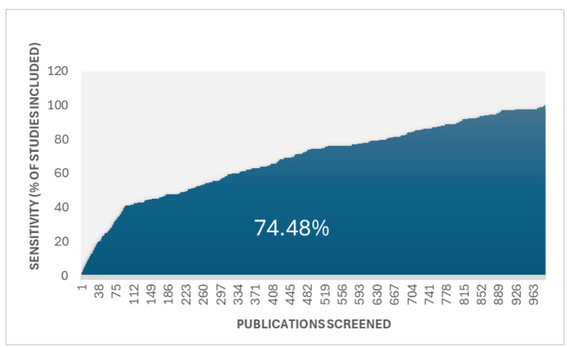
Figure 3. Optimal vs Random screening progress

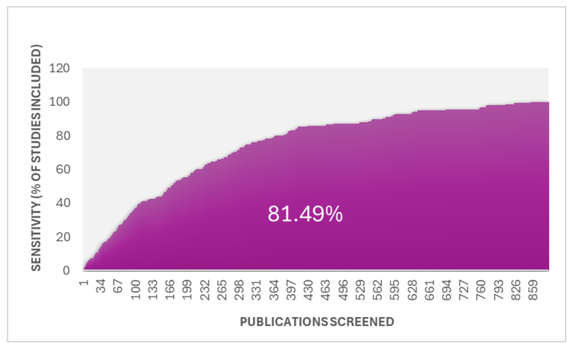


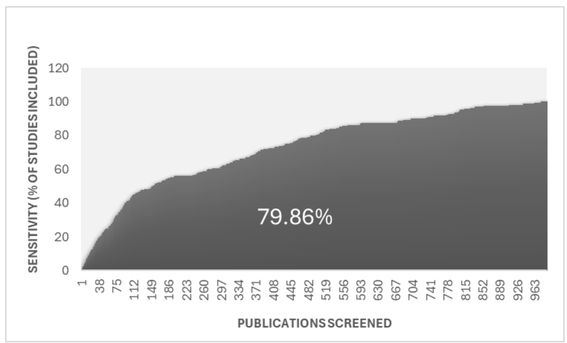
RESULTS

- 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.









CONCLUSIONS

- 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).
- 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

Abstract Title

Authors

Author Affiliations

Abstract No.

INTRODUCTION

- A brief background (4-5 bullet points)

OBJECTIVES

- The objective of the study (mainly one bullet point)

METHODS

- A detailed description of study methodology (good to provide a pictorial presentation)
- If it is a targeted review or a systematic review, the databases searched, dates, and an associated figure of SLR/TLR methodology

RESULTS

- Results

CONCLUSIONS

- Conclusions

REFERENCES

- The reference list generated using Endnote or Zotero
- If the references are more than 10 or 15, a note stating "References available upon request"

DISCLOSURES

XX and YY are employees of Amaris Consulting, which received professional fees from ZZ Pharmaceutical Development and Commercialization, Inc. for the study and has also received fees for projects outside the present study. BB is an employee of ZZ Pharmaceutical Development and Commercialization, Inc.