Paper ML05

Revolutionizing Literature Reviews with a GPT-Assisted Tool: Enhanced Search, Extraction, and Visualization

Martin Schärer, Hoffman La Roche, Basel, Switzerland
Yingjia Chen, Genentech, South San Francisco, USA

ABSTRACT

Conducting literature reviews in drug safety and pharmacovigilance is resource-intensive, requiring manual search, selection, and data extraction, which can lead to inefficiencies and subjective variations. This study presents a GPT-assisted tool that automates key aspects of the process, improving efficiency and reliability through automated data extraction, summarization, and visualization.

The tool, implemented in Python as a Streamlit application using the GPT-4o model from OpenAI, enables customized literature queries tailored to specific research questions in any language. It leverages GPT to categorize papers into "included" and "excluded" groups based on predefined research criteria. In this study, we focus on papers discussing drug-event associations for immune checkpoint inhibitors (ICIs) and immune-related hepatitis. Papers positively identified for inclusion undergo further analysis to extract key measures of association (e.g., ROR, RR, HR). The extracted data are then compiled into forest plots suitable for use in meta-analysis.

Preliminary testing demonstrates the tool's effectiveness in streamlining literature review workflows, significantly reducing the time spent on manual tasks. The GPT-assisted categorization and data extraction features exhibited a high degree of accuracy, with strong concordance with expert reviewers. However, careful prompt design is essential to achieving optimal outcomes.

INTRODUCTION

In pharmacovigilance and drug safety, literature reviews are an integral part of detecting drug safety signals and assessing potential causal relationships between drug exposure and adverse events (AEs). The process typically begins with a request from a safety scientist to the literature review team, who performs a comprehensive search using specialized literature databases such as Clarivate Dialog. These searches, based on predefined keywords, often yield a substantial volume of publications. Due to the large number of results, this initial dataset is outsourced to a contract research organization (CRO) for further refinement.

At the CRO, a dedicated reviewer reviews the full text of each publication to assess its relevance and identify studies containing valuable information. The selected publications are then summarized and returned to the drug safety team in the form of a compiled sheet. Upon receipt, a drug safety scientist within the department carefully reviews and finalizes the list of selected papers, extracting key information necessary for decision-making related to the research question.

Despite the importance of this process, current practices face several limitations:

1. Resource-intensive: The manual nature of searching, reviewing, extracting, and summarizing information requires considerable time and effort, often involving multiple teams and external CROs.
2. Subjective variation: The outcome of literature reviews may vary depending on the expertise and judgment of individual reviewers, introducing inconsistencies.
3. Massive information: The current process lacks an effective way to visualize the vast amount of information from the selected articles, which can hinder decision-making and limit clear communication with stakeholders.

To address these challenges, this study introduces a GPT-assisted tool designed to automate key aspects of the literature review process. By specifying search criteria upfront—focused on a particular drug exposure and AE—the tool enables the reuse of the same literature pool for multiple research questions, eliminating the need to restart the process with each new inquiry. Articles are systematically processed, with full documentation provided for expert review. Additionally, the tool generates visual summaries, such as forest plots, offering an integrated view of the data across all selected studies. This not only enhances working efficiency and consistency but also improves precision and communication with stakeholders, resulting in a more reliable and streamlined literature review process.

In this study, we use papers discussing drug-event associations for immune checkpoint inhibitors (ICIs) and immune-related hepatitis as an illustrative example.

MATERIALS AND METHODS

TECHNOLOGY STACK

The application discussed in this paper uses several technologies and services outlined below.

GPT-4o Model

GPT-4o ("omni") is currently OpenAI's most advanced GPT model. It is multimodal and accepts text and image inputs, although image processing has not been used in the present solution.

Model	Description	Context Window	Max output tokens	Training data
gpt-4o	OpenAI's high intelligence flagship model for complex, multi-step tasks	128,000 tokens	16,384 tokens	Up to Oct 2023

Table 1. Information regarding the AI model used.

Cost as of October 2024:
$2.5 / 1M input tokens
$10 / 1M output tokens

Python and Streamlit

The development of the GPT-assisted literature review tool was implemented using Python 3.9 and Streamlit, chosen due to their combination of flexibility, simplicity, and cost-effectiveness. Python, widely regarded for its robust ecosystem of scientific libraries and frameworks, offers extensive support for machine learning and natural language processing (NLP), making it an ideal choice for integrating the GPT-4o model.

Streamlit, a Python-based framework, was employed for building the application's user interface due to its minimalistic design, ease of use, and rapid prototyping capabilities. Streamlit allows the seamless creation of interactive web applications with minimal code, which is particularly useful for teams focused on data exploration and visualization. Its integration with Python enables developers to focus on writing the core logic while avoiding the complexities of full-stack web development.

By utilizing Streamlit, we also leveraged open-source technology to reduce software licensing costs, aligning with the organization's goal of using free, community-supported tools wherever feasible. Furthermore, our organization provides rsconnect infrastructure, allowing the application to be easily deployed and hosted internally, thus providing secure, enterprise-level deployment without incurring additional costs for external hosting services. This infrastructure not only supports the Streamlit-based application but also ensures that the tool is accessible to the team in a secure and scalable environment.

APIs and Integration

Europe PMC Web Service

To retrieve relevant publications for further analysis, the tool leverages the Europe PMC RESTful Web Service, which provides access to over 33 million publications from multiple sources, including PubMed, Agricola, the European Patent Office (EPO), and the National Institute for Clinical Excellence (NICE). The service allows access to 4.5 million full-text articles and 1.8 million open-access articles, as well as citation networks and reference lists for more than 12.5 million publications.

The tool queries the Europe PMC database to gather both metadata and full-text articles using structured queries that combine specific drug exposure and adverse outcome terms. The query string enables the retrieval of detailed records with customizable parameters, such as publication type and date limits, ensuring that the search is focused on the most relevant research for further analysis.

For example, the following query string 1 was used to retrieve publication data based on predefined drug-event associations:

```
https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=(({exposure}) AND
({outcome})){publication_type_limit}{date_limit}&format=xml&pageSize=550&resultType=co
re
```

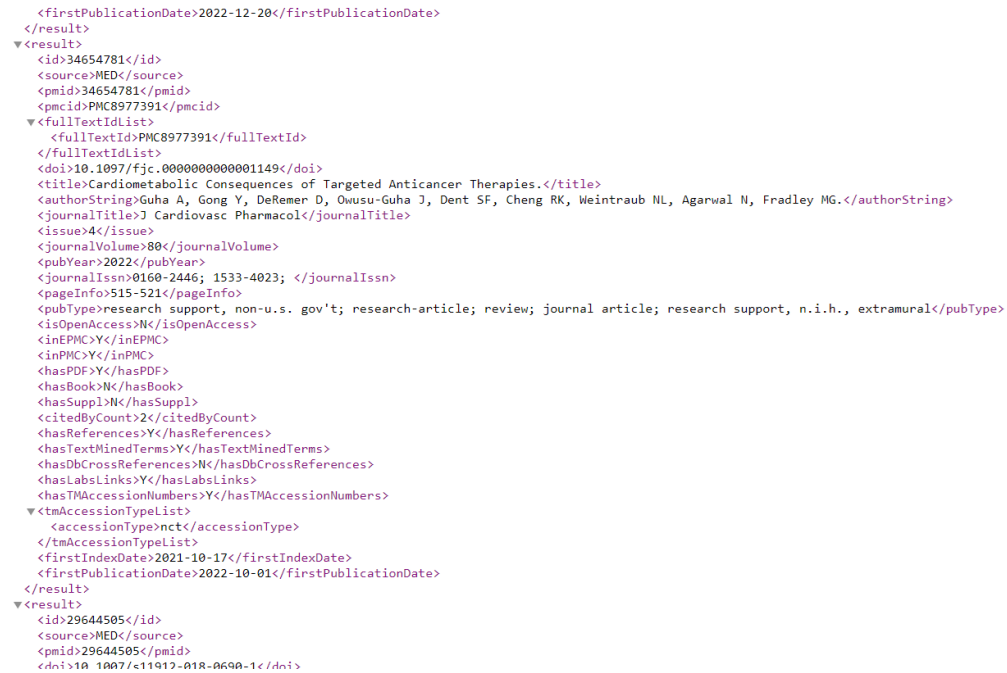

Fig. 1. Example output of query string 1.

Based on the pmid (PubMed Identifier) or pmcid (PubMed Central Identifier for articles available in full-text form) of the identified papers further information such as the abstract and full text is retrieved:

Query string 2 is executed for each publication using its pmid to retrieve the abstracts:

```
https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=(EXT_ID:{pmid})&format=xml&resulttype=core
```

Query string 3 is executed for each publication using its pmcid to retrieve the full text article in XML format.

```
https://www.ebi.ac.uk/europepmc/webservices/rest/{pmcid}/fullTextXML
```

The full text XMLs are then subsequently directly used in prompts to query information about the paper using gpt-4o. Some articles found using the initial query (query string 1) do not have a pmcid and no full-text XML. In such cases full-text information is downloaded from the publisher's website if possible (often blocked) or from the corresponding full-text pdfs if available.

Azure OpenAI

Azure OpenAI Service provides REST API access to OpenAI's language models such as GPT-4o. Parameters used for prompting are:

Parameter	**Value**
API version	2023-10-01-preview
Model	gpt-4o
Temperature	0 - 0.7 depending on prompt
Top P	0 - 0.95 depending on prompt
Max tokens	500- 4096 depending on prompt
Frequency penalty	0
Presence penalty	0

Table 2. Parameters used for the AI model

Example python code:

```
from openai import AzureOpenAI
```

```
api_key = os.environ.get("OPENAI_API_KEY")
azure_endpoint = os.environ.get("AZURE_OPENAI_ENDPOINT")

client = AzureOpenAI(
            api_key=api_key,
            api_version="2023-10-01-preview",
            azure_endpoint=azure_endpoint
        )

response = client.chat.completions.create(
                model='gpt-4o',
                messages=messages,
                temperature=0.7,
                max_tokens=4096,
                top_p=0.95,
                frequency_penalty=0,
                presence_penalty=0,
                stop=None

            )
```

WORKFLOW OVERVIEW

The application that we developed (MetaBooster) is a GPT-assisted tool designed to streamline the literature review and meta-analysis process, offering a comprehensive solution for the identification, categorization, and analysis of scientific publications. The application facilitates an efficient workflow that allows users to conduct broad initial searches, refine the results, and perform advanced analysis on drug-event associations.

Initial Search and Query Definition:

The workflow begins with the user entering parameters to define a broad search of the literature. In our example, the search focuses on an exposure to a class of drugs (such as immune checkpoint inhibitors) and an outcome involving a group of adverse events (e.g., immune-related hepatitis). Users can apply further filters, such as limiting the publication type (e.g., clinical trials, reviews) and setting a date range to refine the search scope. Once these parameters are set, MetaBooster queries the literature databases, retrieving a comprehensive list of publications matching the criteria. Before retrieving the data the user can see how many search hits were found.

Grid Display and Metadata Presentation:

The search results are presented in a dynamic grid that displays essential metadata for each publication. This metadata includes the article title, authors, abstract, a link to the full-text article, and the full text itself when available. This grid serves as the primary interface for users to review the identified publications. In cases where the automatic extraction of the full text was not successful, users have the option to manually edit or input the full text directly into the grid. If the extracted full text exceeds the token limit of the AI model (in our case 128K tokens) a warning is displayed and the user can manually shorten the full-text.

Curation of the Initial Publication List:

Users are given the ability to manually curate the list of publications by excluding irrelevant papers, ensuring that only relevant studies are considered for further analysis. This step is critical in refining the dataset before proceeding with more detailed examination. Manual adjustments can be made directly in the grid, either excluding papers entirely or updating specific fields, such as the full-text content.

Scientific Query and Categorization:

Once the initial list is curated, users can apply a scientific query to further classify the papers. For example, the user might ask, "Does this paper study the association between drug A and event B?" This query is processed by MetaBooster's AI engine, which applies the query to the full-text article of each publication. Using GPT, the tool categorizes the papers into included or excluded groups based on the relevance of the content to the scientific query. The results of this categorization, including the reason for inclusion or exclusion, are displayed in a new grid alongside the metadata.

Final Curation and Adjustments:

After the AI-driven categorization, users have the flexibility to further curate the lists manually, moving papers between the included and excluded groups based on their judgment. This step ensures that the final set of papers reflects both the AI's categorization and any additional expert oversight.

Extraction of Measures for Meta-Analysis:

Once the list of included papers is finalized, MetaBooster retrieves relevant reporting ratios and measures of association from each study, which are essential for meta-analysis. These measures include statistics such as the reporting odds ratio (ROR), relative risk (RR), and hazard ratio (HR). The extracted data are displayed in a new grid, alongside other critical information, such as the source data/database used, the time period studied, the disease population, the sample size, and the type of measure of association along with its explanation.

This section is generated using 6 different prompts each applied to all selected publications.

Prompt	**Prompt text**	**GUI element**
1. Source data	In the article above, first describe the source of the dataset that was used in this study. Secondly put the name of the dataset or database in the tags <Answer></Answer>.	Column "Source Data"
2. Disease population	In the article above, first describe the disease population that was used in this study. Secondly put a short description of the disease population between the tags <Answer></Answer>	Column "Disease Population"
3. Time Span	In the article above, what is the time span of the included observations? Put the time span in the tags <Answer></Answer>	Column "Time Span"
4. Sample Size	In the article above, what is the sample size? Put the number in the tags <Answer></Answer>	Column "Sample Size"
5. Measure of association	In the article above, First describe the measure of association that was used in this study. Then do the following: 1) In tags <MA></MA> put just the term for the measure of association, e.g. RR, OR etc. 2) In tags <EX></EX> put a short explanation of the measure of association used 3) In tags <IN></IN> put an interpretation of the used measure of association	Columns "Measure of Association" "Explanation of Measure of Association" "Interpretation of Measure of Association"
6. Actual values	In the article above, can you find the ratios (RR, ROR, HR, OR) and their 95% confidence interval for a drug and adverse event? Here is an example: 1) ROR for Aspirin vs Advil and Fever is 3.5 with 95% CI (2.6, 9.8). 2) ROR for Aspirin vs Tylenol and Headache is 4.1 with 95% CI (4.0, 4.2) 3) HR for Paracetamol vs Aspirin and Anaemia is 5.5 with 95% CI (4.1, 5.2) In this case write the answer as "<Answer>[ROR] Aspirin vs Advil, Fever:	Column "Measure of Association with 95% CI"

	3.5 (2.6, 9.8) | [ROR] Aspirin vs Tylenol, Headache: 4.1 (4.0, 4.2) | [HR] Paracetamol vs Aspirin, Anaemia: 5.5 (4.1, 5.2)</Answer>" Now put the answer for the paper above in this format: <Answer></Answer>	

Table 3. Prompts used for extraction of key data for meta-analysis

Forest Plot Generation and Saving:

After the extraction of relevant data, users can generate forest plots to visually represent the measures of association, which are a standard tool in meta-analysis. The forest plots can be customized based on the extracted ratios, and the Python code used to generate the graph can be downloaded for further customization or reproducibility. Additionally, at any stage of the workflow, the entire analysis, including curated lists and extracted data, can be saved for future reference or further updates.

To extract the data used in the python code for the forest plots the following algorithm was used:

For each publication with extracted data compile a prompt that contains the authors' names, the publication year and the measures of association. Afterwards the configurable text is added: "

```
For each of these treatment-adverse event combinations determine whether they involve
an immune checkpoint inhibitor (ICI) as well as a hepatic adverse event. Show your
reasoning and then make a final selection, repeating all selected rows after 'Final
Selection:
```

An example prompt would look like:

```
Authors: Lasagna A, Sacchi P.

Publication Year: 2024

Measure of association with 95% CI Data: [IC] ICIs, Liver failure: 1.24 (1.06, 1.42)
[ROR] ICI combination therapy vs ICI monotherapy, Hepatotoxicity: 1.31 (1.80, 2.31)
[HR] Early gastroenterology/hepatology consultation vs no consultation, ALT
normalization: 1.89 (1.12, 3.19)
[HR] Early gastroenterology/hepatology consultation vs no consultation, ALT improvement
to ≤100 U/L: 1.72 (1.04, 2.84)
[HR] Chronic proton pump inhibitor (PPI) use vs no use, GI irAES including IMH: 13.22
(3.11, 56.10)
[HR] Chronic proton pump inhibitor (PPI) use vs no use (after weighing for the
propensity score), GI irAES including IMH: 15.13 (3.22, 71.03)

For each of these treatment-adverse event combinations determine whether they involve
an immune checkpoint inhibitor (ICI)
as well as a hepatic adverse event. Show your reasoning and then make a final
selection, repeating all selected rows after 'Final Selection:'
```

These prompts are then individually run for each selected publication within one session, so that the final summarization prompt has access to all the answers:

```
From the text above extract the data in the form shown in this example. Use first
authors, et al. Leave out rows containing empty values:
[
("Regimen 1,  Adverse Event 1, [RR]", "Zanuso V, et al.", 2023, 0.66, 0.52, 0.85),
    ("Regimen 2,  Adverse Event 1, [HR]", "Pirozzi A, et al.", 2023, 0.65, 0.53, 0.81)
    ...
]
```

The response of the AI model is a list of values that can be directly used in the python forest plot code as a variable definition.

RESULTS

In this section, we present the key outcomes from the implementation and preliminary testing of MetaBooster. The results are organized around the main functions of the tool: search and retrieval efficiency, AI-driven categorization accuracy, extraction of key measures for meta-analysis, and the generation of forest plots.

Initial search and retrieval

MetaBooster's ability to conduct broad literature searches was evaluated using a predefined query targeting studies on immune checkpoint inhibitors (ICIs) and immune-related hepatitis. To limit the search results "atezolizumab" was used instead of ICIs. The search, leveraging the **Europe PMC Web Service**, retrieved **154 publications** with an average retrieval time of 1-2 seconds per publication. No further limits on publication type or time period were made.

The retrieved publications were displayed in MetaBooster's user interface, where key metadata such as article titles, authors, abstracts, links to full-text articles, and the full-text content itself were successfully extracted and presented. Out of the 154 publications, 120 had the full text automatically extracted from the XML full texts, 33 had to use the publication full text pdf and one extraction was unsuccessful and had to resort to the abstract text only. Manual adjustments were made in the user interface and a final selection of 137 publications were used for further analysis.

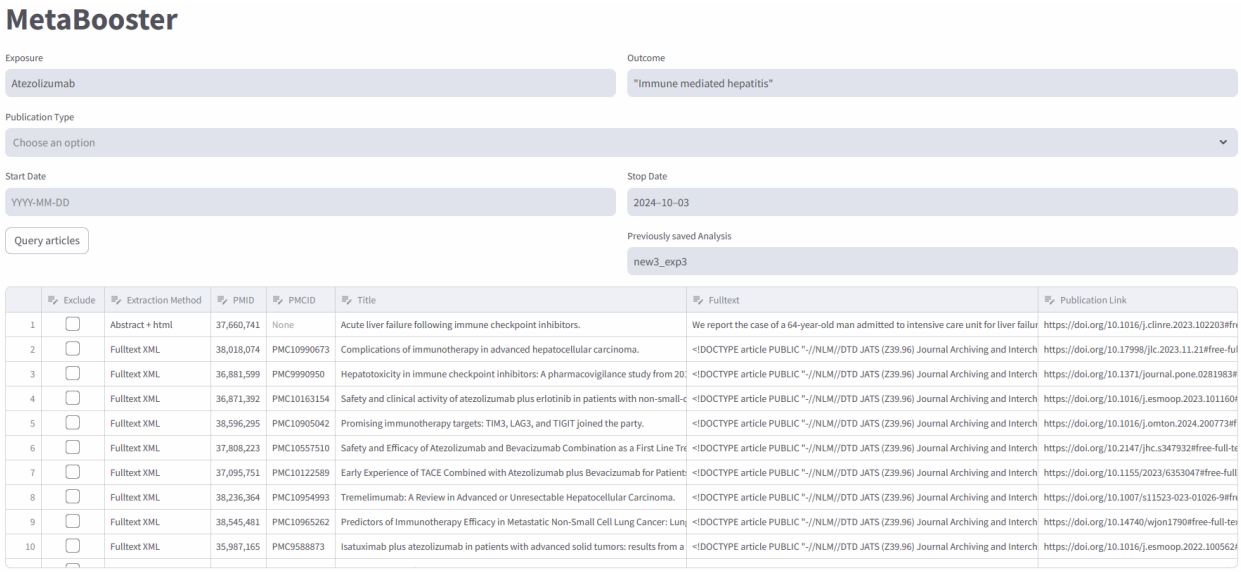

Fig. 2. List of publications retrieved using the initial query.

AI-Driven Categorization Accuracy

MetaBooster's AI-driven categorization feature was tested by applying a scientific query: *"Analyzing this paper, is it possible to extract ratio values (e.g., RR, ROR, HR, etc.) for either individual ICIs (such as Atezolizumab) or for groups of ICIs (such as PD-L1 inhibitors, PD-1 inhibitors, CTLA-4 inhibitors) in the context of "Immune-mediated hepatitis" where different treatment regimens are compared?"* to the curated list of 137 publications. Initially the prompt was significantly simpler but to get the best sensitivity and specificity some testing had to be made to find the most suitable prompt.

The GPT-4o model was tasked with categorizing papers into **included** or **excluded** groups based on the above question. Out of the 137 papers, **6 were included** as relevant, and **131 were excluded**. The categorization process took **7:42 minutes**, significantly reducing the time required for manual screening. Papers were normally around 10,000 to 25,000 tokens in length and API costs incurred for the whole batch were **$12.81.**

In this semi-automated workflow, the machine classifies the study data, while a domain expert oversees the results. Rather than manually reviewing every detail, the expert conducts periodic checks to validate the machine-generated output. When anomalies or unexpected results are observed, the expert refers back to the original raw papers to verify the data and ensure accuracy. This approach combines the efficiency of automation with the critical oversight of domain expertise, allowing for targeted, in-depth review when necessary while maintaining overall process efficiency.

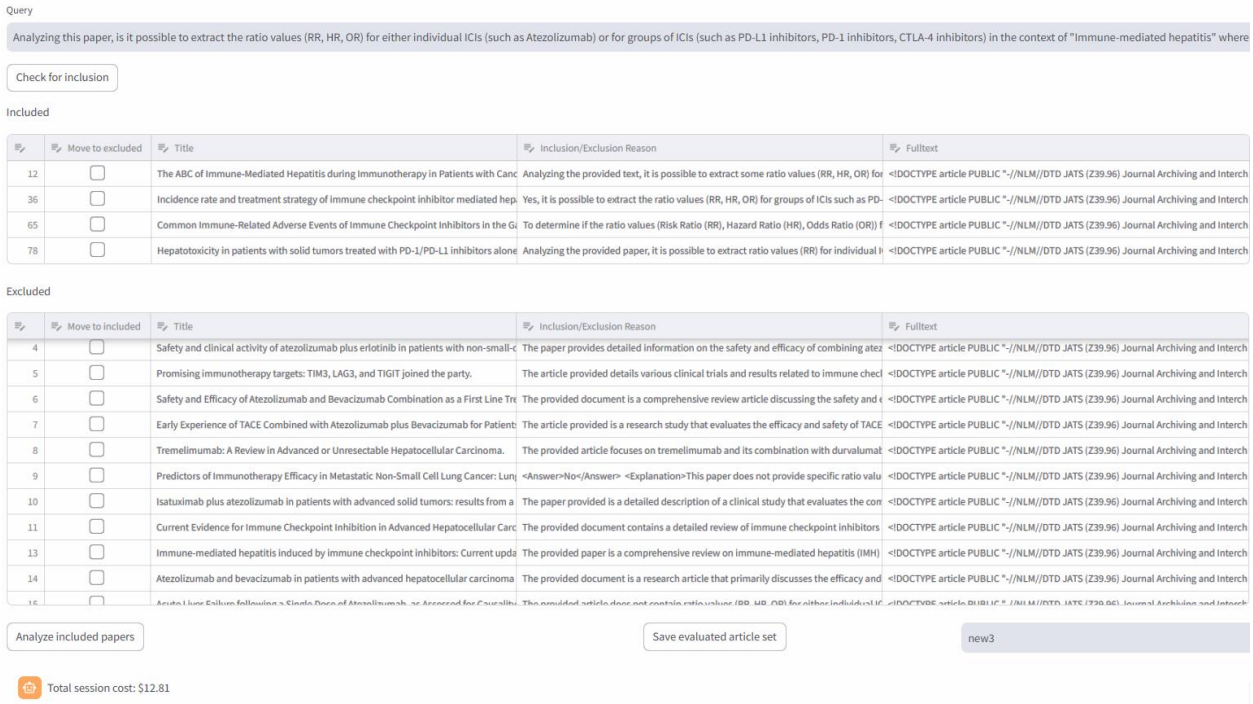

Fig. 3. Selection of included and excluded publications.

Extraction of Key Measures for Meta-Analysis

Once the final list of 6 included papers was established, MetaBooster proceeded to extract key measures of association from each paper. These measures included risk ratios (RR), reporting odds ratios (ROR), and odds ratios (OR), all essential for conducting meta-analyses.

The extracted data were displayed in a new grid, along with additional information such as the source of the data, the disease population studied, the sample size, and the time period analyzed. The ability to retrieve these critical data points substantially saved the time typically required for manual extraction.

Fig. 4. Extracted parameters for use in forest plots

Example Extraction:

Title	Hepatotoxicity in patients with solid tumors treated with PD-1/PD-L1 inhibitors alone, PD-1/PD-L1 inhibitors plus chemotherapy, or chemotherapy alone: systematic review and meta-analysis.
Source Data	The dataset used in this study was sourced from the PubMed and Embase databases.
Disease Population	The disease population used in this study consisted of patients with solid tumors. The study included patients with non-small cell lung cancer (NSCLC), melanoma, carcinoma of the head and neck, small cell lung cancer, gastro-esophageal junction cancer, urothelial carcinoma, and breast cancer.
Time Span	2 to 12 weeks following initiation of ICI treatment
Sample Size	11634

Measure of Association	RR
Measure of Association with 95% CI	[RR] PD-1/PD-L1 inhibitor vs Chemotherapy, High-grade elevated AST: 2.13 (1.04, 4.36) [RR] Pembrolizumab vs Chemotherapy, All-grade Hepatitis: 5.85 (1.85, 18.46) [RR] Pembrolizumab vs Chemotherapy, High-grade Hepatitis: 5.66 (1.58, 20.27) [RR] PD-1/PD-L1 inhibitor plus Chemotherapy vs Chemotherapy, All-grade Hepatitis: 2.14 (1.29, 3.55) [RR] PD-1/PD-L1 inhibitor plus Chemotherapy vs Chemotherapy, High-grade Hepatitis: 5.24 (1.89, 14.52) [RR] Pembrolizumab plus Chemotherapy vs Chemotherapy, High-grade Hepatitis: 7.31 (0.98, 54.44) [RR] Atezolizumab plus Chemotherapy vs Chemotherapy, High-grade Hepatitis: 4.53 (1.39, 14.76) [RR] Pembrolizumab plus Chemotherapy vs Chemotherapy, All-grade Hepatitis: 7.89 (1.05, 59.07) [RR] Atezolizumab plus Chemotherapy vs Chemotherapy, All-grade Hepatitis: 1.82 (1.07, 3.09)
Explanation of Measure of Association	Relative Risk (RR) is a measure used in epidemiological studies that compares the risk of a certain event occurring in one group with the risk of the same event occurring in another group.
Interpretation of Measure of Association	In this study, the RR was used to compare the risk of hepatotoxicity in patients with solid tumors who received a PD-1/PD-L1 inhibitor alone, a PD-1/PD-L1 inhibitor plus chemotherapy, or chemotherapy alone. An RR greater than 1 indicates a higher risk of hepatotoxicity in the treatment group compared to the control group, while an RR less than 1 indicates a lower risk.

Table 4. Extraction of parameters for one selected publication.

Forest Plot Generation and Visualization

After extracting the relevant reporting ratios and measures of association, MetaBooster generated **a plot** to visually represent the meta-analysis results. Careful fine tuning of the methodology to limit only to relevant data points had to be done prior to plotting the results (see Method section). The tool produced forest plots based on the extracted data, which included measures of heterogeneity and confidence intervals.

In addition, the **Python code** used to generate these forest plots was automatically generated and made available for download, allowing users to further customize or reproduce the analysis. Users were also able to save the entire analysis at any point, providing flexibility for iterative analysis and reporting.

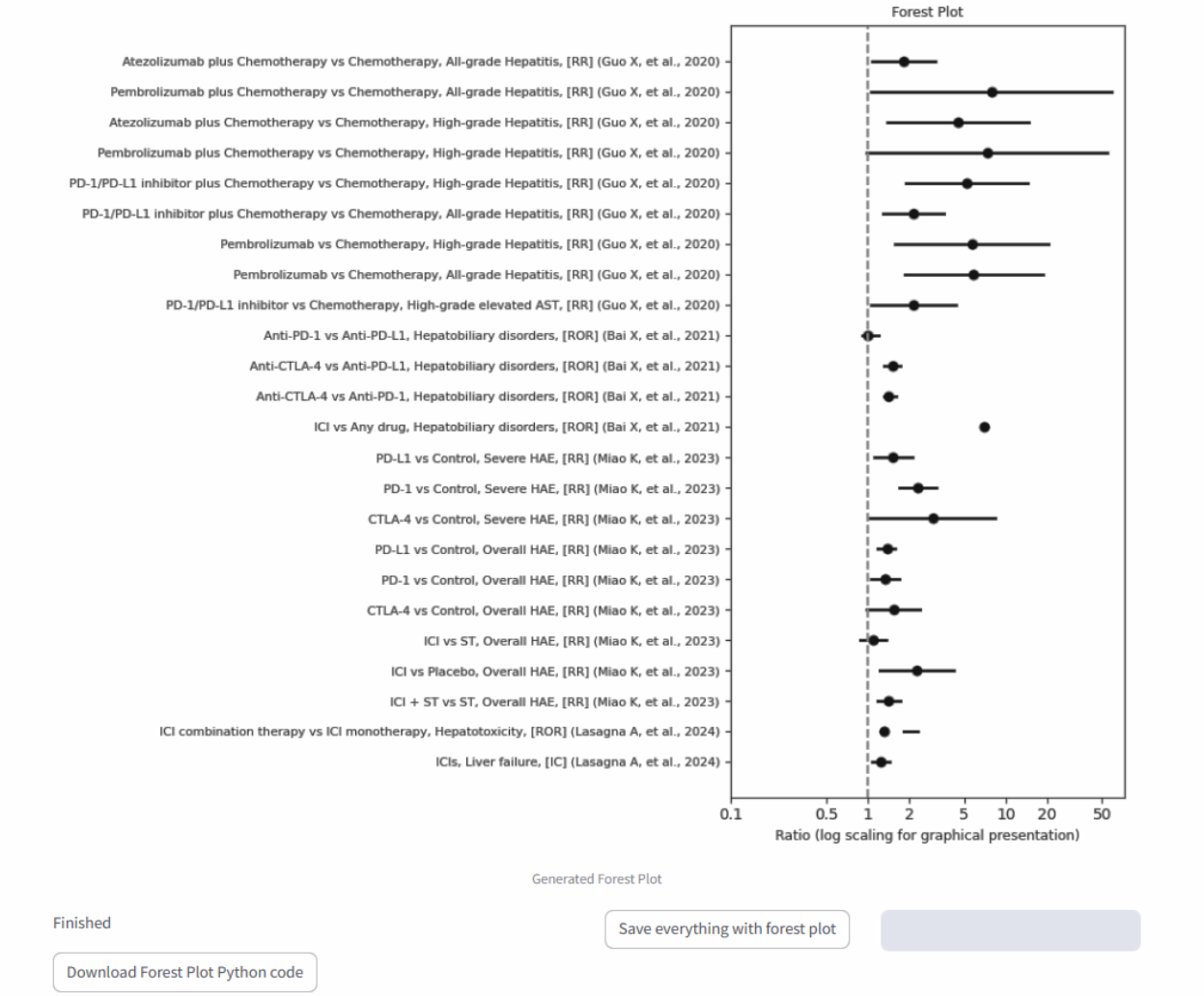

Fig. 5. Generated forest plot.

Interestingly the confidence interval outside the data point in the second last line is indeed specified like this in the original paper:

> grade hepatoxicity = 1.52 vs. 0.48) [15]. The incidence of liver toxicity with anti-CTLA-4 monotherapy ranges from 2% to 15%, while with anti-PD-1 and anti-PD-L1 it ranges from 0% to 3% and 0% to 6%, respectively [16]. ICI combination therapy showed a greater risk of hepatotoxicity compared with ICI monotherapy (IC = 0.63, 95% CI, 0.99–2.04; ROR = 1.31 [95% CI, 1.80–2.31]) [14]. In the phase III clinical trial Checkmate 067, patients with advanced melanoma treated with ipilimumab plus nivolumab presented a higher alanine aminotransferase (ALT) increase (19%) when compared to those receiving ipilimumab (4%) or nivolumab (4%), while the increase in

Thus, the application also visualizes and highlights potential mistakes in the source material.

Impact on Workflow Efficiency

Preliminary testing of MetaBooster demonstrated significant improvements in workflow efficiency and reduced cost.

CONCLUSION

The present study developed a workflow that streamlines literature searches, manual screening, information extraction, and data presentation for meta-analysis, utilizing a GPT-4o-assisted tool. Preliminary results indicate that this approach offers faster and more cost-effective outcomes compared to traditional methods that rely on literature review teams and contract research organizations (CROs). An actual literature search conducted within our company's drug safety department served as a benchmark, revealing an overlap in initial pre-selection results, with some discrepancies. These differences can often be attributed to the inherent interpretative nature of scientific queries. Notably, our tool

demonstrated potential in identifying publications that might otherwise be overlooked during manual reviews. However, further quantification is needed to more accurately assess its efficacy.

A significant limitation of this approach lies in the time spent refining prompts and optimizing search strategies, often requiring multiple iterations that added to the time and costs. The goal is to make these initial efforts reusable for similar future tasks, thus enhancing long-term efficiency. While the tool can, in principle, be adapted for various literature analysis needs, the current implementation is specifically tailored to drug therapy-adverse event combinations and their measures of association.

Prompt design emerged as a crucial factor in achieving consistent results, yet it presented challenges. Prompts that worked well in one scenario sometimes failed in others, with adjustments to model parameters like temperature and top_p offering only marginal improvements. A key issue was that overly complex prompts often led to imprecise outputs. For simpler tasks, such as extracting a few descriptive values, a single prompt could efficiently retrieve multiple details like the type of measure, its explanation, and interpretation. However, more complex prompts, such as those used for summarization and formatting the actual values (see table 4, Measure of Association with 95% CI) required highly specific instructions and examples to ensure accuracy.

The initial prompt for including or excluding relevant publications needed careful fine-tuning to balance specificity and sensitivity. Without specifying that measures of association (e.g., RR, ROR, HR) should be present, the model would sometimes include irrelevant publications, which then complicated subsequent data extraction. Therefore, precise filtering was essential from the beginning. Additionally, during the extraction phase for forest plot generation, the model occasionally included unrelated data points, which necessitated a multi-step approach to achieve better consistency. This included individual prompts for each publication, explanations for data point inclusion/exclusion, and a final summarization into Python code for generating plots.

The following insights regarding prompt design emerged throughout this study:

1. **Initial Classification**: Optimal results in classifying relevant publications were achieved by using highly detailed prompts. These prompts include a comprehensive description of the research problem, accounting for edge cases while excluding irrelevant papers. Furthermore, specifying the presence of key values for extraction (e.g., RR, ROR, HR) in the initial prompt helps filter out studies that, although related to the topic, do not contain the necessary data for subsequent analyses.
2. **Data Extraction from Publications**:
    - Simple summarization tasks, such as identifying databases or study populations, can be effectively managed using a single prompt or even a prompt that extracts multiple values simultaneously.
    - Extracting and formatting values for analysis, such as those used in forest plots, proved more complex. The most effective strategy involved:
        1. Using precise prompts with examples to extract formatted values (e.g., "[RR] Pembrolizumab plus Chemotherapy vs Chemotherapy, All-grade Hepatitis: 7.89 (1.05, 59.07)") without seeking additional context at this stage.
        2. Dividing the extracted values into manageable subsets (e.g., handling all values for each publication individually) and prompting the model to assess the relevance of each value to the scientific question. Requesting explanations for each classification decision improved the accuracy of this step.
        3. Summarizing the refined values from step 2 and formatting them for integration into Python code for forest plot generation.

This approach to prompt design not only enhanced the precision of data extraction but also contributed to a more streamlined and reproducible analysis process.

The transition from GPT-4-32k to GPT-4o during development significantly improved both performance and cost efficiency. The architecture of the application allows for seamless integration of new AI models, suggesting that the tool's performance will continue to improve as OpenAI releases more advanced models. However, each model update requires prompt adaptation to maintain accuracy.

Overall, while challenges remain in prompt optimization and complex query handling, this study demonstrates that a GPT-assisted workflow has the potential to substantially improve the speed, efficiency, and comprehensiveness of literature reviews in drug safety and meta-analysis. When using this tool, drug safety scientists—and potentially other users beyond this field—should be mindful that while it significantly streamlines literature reviews and enhances efficiency in data extraction, summarization, and communication, formal decision-making still heavily relies on their domain expertise. The tool assists in automating routine tasks, but the interpretation of results and the integration of both qualitative and quantitative safety research remain crucial. Users should view the tool as an aid to improve workflow, not as a replacement for the critical scientific judgment necessary for comprehensive and reliable safety evaluations.

REFERENCES

Kiciman, E., Ness, R., Sharma, A., & Tan, C. (2024). *Causal Reasoning and Large Language Models: Opening a New Frontier for Causality*. Transactions on Machine Learning Research. https://openreview.net/forum?id=mqoxLkX210

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our main collaborators, Svetlana Lyalina, Seid Hamzic, and Bamboo Lin, for their invaluable expertise and insightful comments. We are also deeply appreciative of Robert Walls, Christian Müller, and Andrew Hudson for their ongoing support and encouragement. Lastly, we extend our thanks to Monika Kaul and Rajesh Ghosh for their valuable contributions throughout the project.

CONTACT INFORMATION

For comments and questions please contact the authors at: martin.schaerer@roche.com, cheny224@gene.com