A no-code method for explorative testing of scientific hypotheses in drug discovery

Nina Truter[1], Raminderpal Singh[2]

Affiliations:
[1]Independent Consultant, South Africa
[2]HitchhikersAI, London, United Kingdom

Abstract

Drug discovery scientists spend their day developing and testing complex hypotheses, taking advantage of data and know-how. To enable this, they build workflows using available tools. In this paper, we introduce a high level workflow for scientists to use. We describe the functions in the workflow, discuss key considerations, and summarise computational tools that can be used. We also include an example walkthrough, which uses ChatGPT to analyze mice study data.

1. Introduction: proposed workflow for hypothesis testing in drug discovery

In drug discovery and biological research, the scientist's workflow often follows a structured and iterative approach to ensure accuracy, reproducibility, and scientific integrity. This article outlines the various stages of such a workflow, from hypothesis generation to data cleaning and interpretation, referencing the project workflow diagram (Figure 1).

At the heart of any research is the scientific question. This question shapes the direction of the entire investigation and helps to focus the analysis on specific hypotheses. Following the definition of the scientific question, it is important to generate hypotheses based on prior research and available datasets. These datasets, which may be either public or proprietary, provide the foundation for all subsequent analysis. Public datasets, while larger, often require careful scrutiny as they may contain noise or inconsistencies in how the data were collected, processed, or labelled that need to be accounted for before beginning analysis. For example, RNA sequencing datasets might have incomplete information on the experimental conditions under which the data was generated, leading to potential misinterpretations if such details are not properly understood.

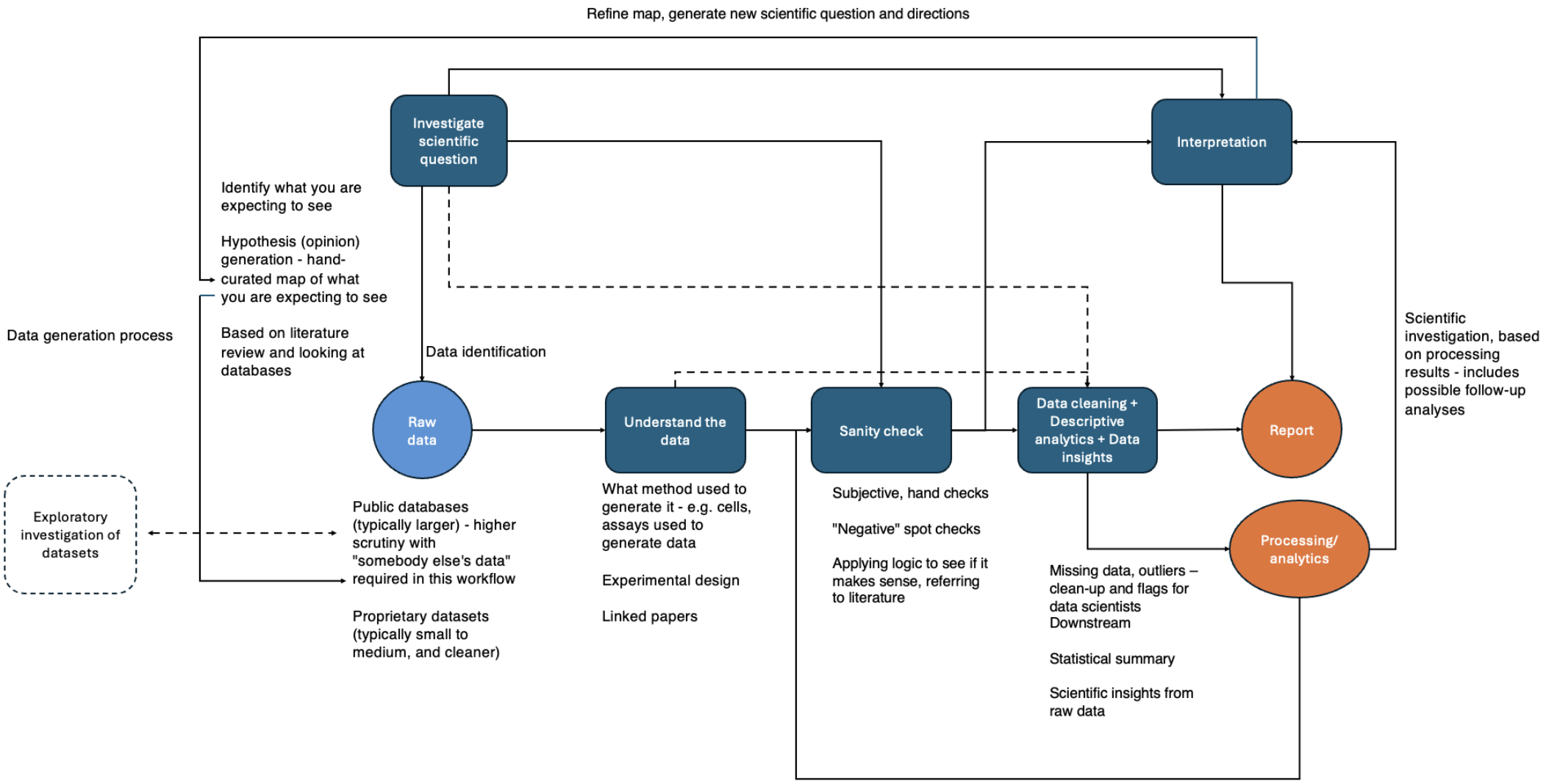


Figure 1: High-level method for early drug discovery

On the other hand, proprietary datasets tend to be smaller and are generated under more controlled conditions, such as those derived from in-house assays or experiments conducted within a lab. These proprietary datasets typically do not require as much validation as public data but still necessitate a solid understanding of their experimental design and data generation methods in order to determine if they are suitable to answer the scientific question.

Once the raw data has been gathered, the next step is to gain a thorough understanding of the data. This includes verifying the experimental design and understanding how the data was generated—whether it was from RNA sequencing, mass spectrometry, or other biological assays. Public datasets, linked papers or supplementary documents often provide essential context about the experimental setup. These linked materials may describe the cell lines used, the specific conditions under which experiments were performed, and any potential limitations of the data. Without this information, the interpretation of the dataset may be speculative or lead to inaccurate conclusions.

The next critical step is the “sanity check” phase, where the integrity of the dataset is assessed. This process involves checking for biological inconsistencies or errors in the data, such as the presence of genes that should not be expressed in certain tissues. For example, while working on diabetes research, one can include proteins involved in bone metabolism in the dataset - unexpected alterations in bone metabolism proteins could indicate that something is wrong with the data, necessitating further investigation before proceeding with analysis.

After the sanity checks are completed, the next stage is the data cleaning phase. In this stage, missing data points, outliers, and other anomalies are handled. *In vivo* datasets, which involve experiments conducted on living organisms, often exhibit natural biological variation and outliers in such datasets may actually represent meaningful variability. For example, in rat studies some control groups might show unexpectedly high values compared to rats from a treated group when investigating a highly variable marker such as cytokine levels. In these cases, removing outliers could erase significant biological information and are often retained in especially smaller datasets. The data cleaning process in *in vivo* studies requires a nuanced approach to balance removing potentially erroneous data and preserving important biological signals.

Once the data has been cleaned, it is then subjected to descriptive analytics. This stage involves generating statistical summaries that reveal patterns, trends, or anomalies in the data. Through methods such as statistical tests, plots, and data visualisation, researchers can begin to interpret the data and form insights. This phase often helps confirm whether the

data aligns with the original hypothesis or if new patterns have emerged that warrant further investigation.

The final step is the generation of a report. This document summarises the data analysis process, including the key findings, statistical analyses, and interpretations. Reports are often shared with collaborators or stakeholders and serve as a record of the investigation. Importantly, the report can also raise new scientific questions, leading to further iterations of the workflow. This iterative nature of scientific investigation often leads to new discoveries as researchers refine their understanding of the data.

2. Important considerations and recommendations

Operating a workflow like the one described in Figure 1 involves several important considerations that impact both the accuracy of the results and the efficiency of the research process. These are described below.

A) Defining the research question

A well-defined research question is the cornerstone of an effective scientific workflow in drug discovery. The more specific your question, the easier it becomes to identify relevant data and design subsequent steps in your workflow. This initial phase often involves an iterative process: refining your question, conducting a literature review, and assessing available data to ensure the right level of specificity and relevance of your research question. AI tools like ChatGPT can help refine your question and provide an overview of the research landscape before you dive into a full literature review.

B) Hypothesis generation

The hypothesis generation process is equally important. Before diving into data analysis, a hypothesis must be developed based on literature reviews and public datasets. The scientific question guides the entire investigation, and without a clear hypothesis, the research could become unfocused and exploratory. Having a well-defined hypothesis allows researchers to assess datasets critically and ensures that their analysis remains grounded in the biological context. Creating a rough map containing the relevant variables that influence the outcome of the scientific question based on literature review and logic can help structure the hypothesis. This map can be used as a “checklist” when assessing whether a dataset contains the necessary variables to answer the research question.

C) Data identification

When searching for public data, tools like Perplexity.ai can aid in the process of identifying relevant databases by for example, asking *"Which database should I use to search for data on the effects of longevity drugs in rodents?"*. While ChatGPT and Claude.ai are useful for general information to questions, Perplexity.ai tends to provide more accurate, "fact based" answers. Google Dataset Search or PubMed's "Associated Data" feature can uncover datasets linked to publications. After identifying a potentially useful dataset, Claude.ai can summarize experimental methods to determine if the dataset is the right fit for your research question. Creating a descriptive spreadsheet to catalog potential datasets, along with a broad description of their contents, helps streamline the selection process. In some cases, combining multiple datasets may be necessary to comprehensively address your research question.

D) Understanding Data

Before diving into analysis, ample time should be spent reviewing the raw data. Browsing through datasets, often in Excel format, can clarify how the data were generated, helping you choose appropriate analytic methods and establish sanity checks. For data types that are less familiar, ChatGPT can be helpful in explaining the experimental method and for establishing potential validation steps. Alternatively, search for review papers or papers using a similar method and understand how it was applied in that context.

Visualization is another powerful tool for data understanding—experimenting with different methods can provide varied perspectives. ChatGPT can also aid in deciding which visualisation options are available and what information each will provide, based on the data and your research question. Additionally, running analyses on both the raw/"uncleaned" and "cleaned" versions of the dataset helps assess the impact of outliers and can guide decisions on whether to include or exclude them.

E) Analyzing and Interpreting Results

When it comes to data analysis, Claude.ai has analytics tools that offer specific methods which can improve the data analytic process. Although ChatGPT is helpful as an initial step to understand results, it should be used as a tool for creating literature review ideas and hypothesis generation, not as a fact-based system. The scientific question should stay the anchor of the interpretive process, together with your understanding of the raw data and output from analytics. Here, it is helpful to toggle between two mindsets - one of a creative scientist, which is useful for creating avenues of exploration and one of a critic when assessing the merit of these avenues.

F) Exploratory investigations and missed opportunities:

Often, datasets are generated for a specific research question, but they may contain additional information that could be useful for answering new or unrelated questions. This is particularly true for large public datasets, where the breadth of data available can sometimes be overwhelming. Researchers may miss opportunities to generate new insights simply because they are focused on their initial question and do not have the resources to explore other possibilities.

Additionally, exploratory analyses can be valuable for identifying new biological markers or hypotheses. For instance, a dataset generated to study protein expression in one context might also reveal valuable information about other biological pathways or processes. However, exploratory investigations can be resource-intensive, both in terms of time and computational power. Researchers need to balance their focused analysis with the potential for broader discoveries.

3. Example tools

A. Data visualization and hypothesis generation tools:

Tools like Miro, a diagram-making tool, are essential for mapping out hypotheses. Miro allows researchers to visually map out the relationships between proteins, genes, or pathways, helping to clarify the expected interactions within the biological system being studied. This kind of visualisation is particularly useful during the hypothesis generation phase, where researchers are still exploring the relationships between different biological components.

ChatGPT is ideal for brainstorming and generating new research ideas. ChatGPT can be used to explore possible pathways or protein interactions by inputting key terms or genes. This tool, while useful for generating ideas, should be used cautiously. It can provide new pathways or hypotheses to investigate but should not replace rigorous literature review or empirical evidence.

B. Data cleaning and descriptive analytics tools:

Excel remains one of the most commonly used tools for data cleaning and descriptive analytics in many research settings. Researchers use Excel for tasks such as sorting data, identifying outliers, and generating basic plots. However, for larger datasets, Excel has its limitations in terms of both scalability and complexity. Tools like R and Python, equipped with

libraries like Pandas for data manipulation and Matplotlib for visualisation, offer more robust solutions for handling larger datasets and performing advanced statistical analyses. Python's Scipy and Statsmodels libraries, for instance, provide advanced tools for hypothesis testing, regression analysis, and other complex statistical procedures that go beyond what Excel can offer. ChatGPT and Claude.ai are useful tools to empower scientists with no coding experience by providing custom-written code for specific analyses and execution of this code. Again, this is not a replacement for rigorous analyses by data scientists, however where data scientists are not available, this allows exploration of the data beyond the capabilities of Excel.

Another powerful tool in the workflow is the KEGG Pathway database, which helps researchers map out how proteins and genes interact within known biological pathways. This is particularly useful during the hypothesis testing phase, as it allows researchers to visualise where their proteins of interest fit into larger biological processes. The KEGG Pathway database provides insights into metabolic pathways, genetic interactions, and disease mechanisms, which are crucial for understanding how a dataset can inform our understanding of complex biological phenomena such as signal transduction, cell proliferation, or immune responses.

Gene ontology databases, such as STRING and Reactome, are additional tools that can be used to understand protein-protein interactions and their involvement in cellular processes. These tools are essential for interpreting the results of data analysis, particularly when the dataset reveals unexpected or novel interactions between proteins that require further investigation.

C. Network and interaction mapping tools:

With increasing complexity in biological datasets, graph-based tools have become essential for visualising and analysing protein-protein interactions and gene networks. Cytoscape, for example, is a widely used software tool for visualising molecular interaction networks and integrating these with gene expression profiles and other data. In research focused on drug discovery, understanding the interactions between multiple proteins or genes is critical for identifying potential drug targets or understanding the mechanisms behind drug resistance.

Network-based approaches are also becoming more prevalent as researchers aim to represent complex biological data in more intuitive ways. By visualising data as networks or graphs, scientists can more easily identify hubs, bottlenecks, or key players in biological processes, allowing them to focus their efforts on the most critical components of a system.

D. Literature and data curation tools:

Data curation is a key part of any workflow, particularly when working with large datasets or integrating data from multiple sources. Databases like GeneCards are useful for obtaining detailed information about genes and their functions. GeneCards offers comprehensive gene-related information, such as pathways, interactions, and diseases associated with each gene. This information is invaluable when generating hypotheses or validating findings, as it provides a deeper understanding of how a particular gene or protein fits into the broader biological context.

In addition to GeneCards, Mendeley or Zotero can be used for managing research papers and references. These tools are particularly useful for researchers, who rely heavily on literature reviews to support their hypotheses and analyses. Properly managing references and associated data ensures that researchers can track their sources efficiently and maintain the integrity of their work.

E. AI and Machine Learning tools:

As datasets in biological research grow in size and complexity, the use of AI and machine learning tools becomes increasingly important. ChatGPT can be used as a brainstorming tool for generating hypotheses or exploring possible pathways. While this tool is still relatively novel in the research community, it represents the growing intersection between AI and drug discovery. ChatGPT can assist by summarising literature, suggesting new angles of inquiry, or even helping to explore large datasets in ways that would be too time-consuming for manual review.

Other machine learning tools, such as TensorFlow or PyTorch, can be used to analyse large datasets and identify patterns that may not be immediately apparent through traditional methods. These tools allow researchers to build predictive models, classify data, or identify novel associations between variables. In drug discovery, machine learning models have been used to predict drug efficacy, optimise compound structures, and even simulate biological systems.

List of tools and databases used in the workflow:

1. KEGG Pathway Database - The KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database provides information on molecular interaction and reaction networks for various biological pathways. https://www.kegg.jp/kegg/pathway.html

2. STRING Database - STRING is a database of known and predicted protein-protein interactions, integrating both physical and functional associations, https://string-db.org

3. Reactome - Reactome is an open-source, curated pathway database that provides insights into biological processes and molecular interactions, https://reactome.org

4. GeneCards - GeneCards is a comprehensive database that provides detailed information on all known and predicted human genes, including functions, pathways, and related diseases, https://www.genecards.org

5. Cytoscape - Cytoscape is a software platform for visualising molecular interaction networks and integrating these networks with gene expression profiles and other data, https://cytoscape.org

6. Mendeley - Mendeley is a reference manager and academic social network that helps researchers organise research papers, collaborate online, and discover the latest scientific research, https://www.mendeley.com

7. Zotero - Zotero is a free, easy-to-use tool to help researchers collect, organize, cite, and share research, https://www.zotero.org

8. TensorFlow - TensorFlow is an open-source platform for machine learning, commonly used for deep learning applications and large dataset analysis, https://www.tensorflow.org

9. PyTorch - PyTorch is an open-source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, https://pytorch.org

4. An example walkthrough

In this walkthrough, ChatGPT was used to generate recommendations for an early clinical trial for aging-related diseases on the dose, participants and potential measurements of efficacy based on results from a primary study on acarbose treated mice [1,2] and a selection of related papers on acarbose.

The prompts and dataset used in this example are available as supplementary material (or can be downloaded from the two download buttons at the bottom of this page). Table 1 shows how the scientific workflow was applied in this example - this table is continuously updated here.

Table 1: Scientific workflow steps and how they were applied to this example

	Workflow "step"	What was done?
1	**Investigate scientific question**	1. Define the scientific question to be answered: "What dose of acarbose should be used to perform early clinical trials on aging-related diseases and which participants should be considered?" 2. Perform a literature review to identify a handful of relevant papers on acarbose in diabetes research (its original treatment indication), acarbose interventions in mice using a combination of Pubmed searches and ChatGPT to identify what type of papers should be looked for/search terms to use in answering the scientific question 3. Read through papers to understand the state of acarbose research and identify the variables/features that should be considered when answering the scientific question, using ChatGPT to summarise papers and highlight these features 4. Create rough map of how the identified variables/features will influence the scientific question and create hypothesis based on these and logic - e.g. from the literature review, we have identified that there are three doses that are used in acarbose research (low, middle, high) and that there are a handful of measures of efficacy in aging-related disease research (lifespan, body weight and functional measurements such as grip strength). The hypothesis is that the middle dose would have the best balance between efficacy based on these measurements and prevention of unwanted effects
2	**Data identification/ Raw data**	One of the papers that was identified linked to a public database/dataset which measured different variables of efficacy after acarbose treatment in male and female mice, which matches the variables/features that were described in the rough map (e.g. lifespan, body weight and functional measurements such as grip strength)
3	**Understand the data**	1. Read through the paper attached to generation of this dataset (https://pmc.ncbi.nlm.nih.gov/articles/PMC6413665/) to understand the design of the experiment/variables that were measured, here the most important variables to take note of was that male and female mice were used, the datasets from the public database are not all linked (performed on different sets of mice), multiple compounds beyond acarbose were measured, experiments were performed across different sites (at different laboratories) 2. Upload each dataset to ChatGPT and asked for an overall description of each dataset and description of each column in the dataset
4	**Sanity check**	Based on understanding the experimental design, determine what observations are expected in the data based on logic, e.g. acarbose works through slowing digestion of carbohydrates, therefore we expect to see a lowering of postprandial blood glucose levels. From the graph in the original paper, we observe a dose-dependent decrease in postprandial blood glucose levels compared to control mice - this matches what is expected and gives confidence in the generated data.
5	**Data cleaning/ Descriptive analytics**	1. Use ChatGPT to re-generate dataset but by excluding mice with "removed" from the status column (determined by description of columns in "understand the data step". Also asked it to remove other compounds that were also studied (Ursolic acid). Hand checked the data with the original dataset to make sure data was not changed. 2. Manual spot checks showed that there were a few mice that only had body weights for week 1, we can decide to keep them in because this is a true representation of the

		average weight of mice during that week, but decided to remove these mice as we are more interested in how the body weight changes - using mice that have more than one data point.
6	**Processing/ analytics and Interpretation**	1. Ask Chatgpt to generate tables that summarise the mean/standard deviation of each of the measurements for untreated vs different doses of acarbose treated male and female mice and provide an interpretation of what each of these would mean for the recommended dose and participants for an early clinical trial: "By displaying data points in tables, compare the female and male mice of the control group to the ACA_lo, ACA_mid, ACA_hi group in terms of the effect on median lifespan, mean body weight, mean body composition, mean fat pads, mean glucose, mean grip strength, mean grip duration, mean rotarod and mean pathology." 2. Ask ChatGPT to compare these results to those from other papers that treated mice with acarbose and asked how this would impact the recommendations 3. Ask ChatGPT to compare the recommendations of a dose to that of papers describing human clinical trials and doses of acarbose that have been used in diabetes research. 4. A new research question was generated from this step as we noticed that female mice showed functional improvement but their increased lifespan was not as large as that observed in male mice. Question: "Why do female mice have different responses to acarbose compared to male mice?"

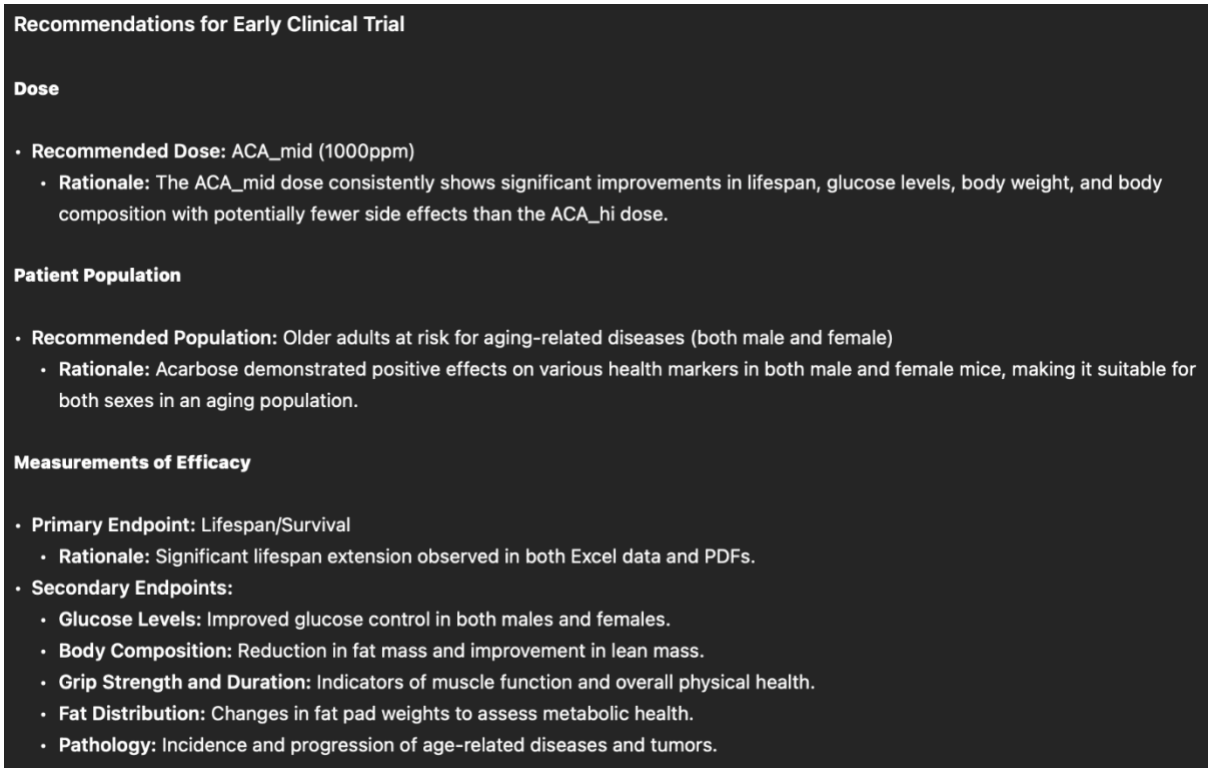
Importantly, commonly accessible LLM systems often share provided inputs, therefore it is recommended not to enter confidential information.

One of the key challenges in using ChatGPT for interpretation is creating prompts to accurately extract information to support recommendations and accurately describing the content of multiple files and papers. To help ChatGPT provide useful insights, there needs to be some ‘prompt engineering’. This is a technical term for best-practices in the way prompts are written. As an example, the first prompt in this example is only to provide background and context to ChatGPT:

> “You are a drug discovery scientist looking to make decisions on dose, participants and measurements when taking an existing diabetes drug into the ageing-related diseases field. You have experimental results from a mouse study that show the effects of acarbose on lifespan, body weight, body composition, fat pads, glucose, grip strength, grip duration, rotarod and pathology. You also have several relevant scientific publications with studies investigating the effects of acarbose on different measurements in mice. You now want to interrogate your study results (which are in Excel files and images) and the publications separately for insights, and then together to get the best set of recommendations for your colleagues who are looking

to perform early clinical trials with acarbose on ageing-related diseases. To do this, you will now process a series of specific user-entered ChatGPT prompts."

The screenshot below shows the results from the last prompt. There are some nuances ChatGPT has not picked up on. For example, in female mice, the lifespan is not extended as much compared to male mice, but their physical measurements are improved. Improved prompts will aid the generation of more nuanced results.



Recommendations for Early Clinical Trial

Dose

- **Recommended Dose:** ACA_mid (1000ppm)
  - **Rationale:** The ACA_mid dose consistently shows significant improvements in lifespan, glucose levels, body weight, and body composition with potentially fewer side effects than the ACA_hi dose.

Patient Population

- **Recommended Population:** Older adults at risk for aging-related diseases (both male and female)
  - **Rationale:** Acarbose demonstrated positive effects on various health markers in both male and female mice, making it suitable for both sexes in an aging population.

Measurements of Efficacy

- **Primary Endpoint:** Lifespan/Survival
  - **Rationale:** Significant lifespan extension observed in both Excel data and PDFs.
- **Secondary Endpoints:**
  - **Glucose Levels:** Improved glucose control in both males and females.
  - **Body Composition:** Reduction in fat mass and improvement in lean mass.
  - **Grip Strength and Duration:** Indicators of muscle function and overall physical health.
  - **Fat Distribution:** Changes in fat pad weights to assess metabolic health.
  - **Pathology:** Incidence and progression of age-related diseases and tumors.

Figure 2: Prompt results.

5. Conclusions

In this paper, we have introduced a high level implementation method for scientists to develop and test complex hypotheses. The method enables scientists to apply data science tools and techniques in a pragmatic effective manner.

6. References

1. Alavez S, *et al*. Acarbose improves health and lifespan in aging HET3 mice. *Aging Cell*. 18(2) (2019 April). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6413665/

2. Harrison DE, *et al*. ITP: Interventions Testing Program: Effects of various treatments on lifespan and related phenotypes in genetically heterogenous mice (UM-HET3) (2004-2023). *Mouse Phenome Database*. Available at: https://phenome.jax.org/projects/ITP1