PharmaSUG 2026 - Paper AI-344

Protocol Analysis, Optimisation and Generation: Artificial Intelligence Enables a Unified View

Sundaresh Sankaran & Sherrine Eid, SAS Institute

ABSTRACT

A successful clinical trial maximizes multiple objectives defined by relevant primary and secondary endpoints, enrolment and retention, statistically significant results and rigorous evaluation. To design a comprehensive, executable and data-driven protocol, inputs are sourced from clinicians, epidemiologists, statisticians, programmers, and compliance officers from a protocol review board to ensure that all requirements are met.

These activities involve significant financial and quality costs. Each step of the way is a potential introduction for quantifiable error. Existing solutions tend to be siloed and only address one task at a time.

We propose a unified approach to minimise errors and inefficiencies in clinical trials through an end-to-end solution for protocol generation which allows seamless handoffs. The solution uses Artificial Intelligence (AI) and Agentic workflows on SAS Viya and Retrieval Agent Manager (RAM) to automate and optimize repeated tasks for data ingestion, search and retrieval, benchmarking, simulation and generation. Agents are capable of reflection and automatically retrieve information, trigger simulation and generate consumable insights. Controlled by humans in the loop, results are standardised and reviewed and enable generation of a draft protocol with maximum first-pass yield. Output drafts are editable and can be collaboratively reviewed prior to finalisation.

This session gives you a recipe for automated and efficient, data-driven protocol design and generation. It suggests a Generative AI framework as an enabler to increase your productivity and unlock time and bandwidth for high-value tasks.

.

INTRODUCTION

Protocol design and draft generation involves many tasks which need to be seamlessly orchestrated for efficiency. Design and generation of a draft protocol require careful attention because it helps ensure success in the actual trial. Protocols are characterized by their importance, length and complexity, and their rigorous scrutiny from a protocol review board at various stages. It is to be expected that this scrutiny and governance may raise questions that lead to rework, for which you should explore technology-based solutions to reduce. This is keeping the principle of ‘first-pass yield’ in mind.

Protocols facilitate clinical trials in achieving their study objectives. These study objectives are usually highlighted as endpoints which are specific outcomes that measure whether an intervention works, is safe, or is both. In addition to specifying a roadmap that defines and plans towards these endpoints, other related concerns involve:

1. Define eligibility criteria which maximises enrollment while maintaining focus
2. Ensure comparability, reduce unwanted redundancy and facilitate comparison and benchmarking with other past studies
3. Ensure traceability with internal and external artifacts such as research documents, past protocols, statistical analysis plans etc.
4. Ensure compliance and enable review by a Protocol Review Board and other internal and external governance teams
5. Allow for participation of all concerned stakeholders such as patients (from experience, safety and need perspective), sponsors, medical writers, statisticians and others

Keeping this in mind, technology needs to enable the activities that go into protocol design. However, technology solutions tend to focus on one part of the process over another. This is understandable given

that many technology solutions are focused on specific parts of the process. For example, tools and solutions that help generate a draft protocol focus on Large Language Models and generation but may lack visibility into the data quality upstream.

You require a unified reference architecture which highlights the key capabilities provided by different components and focuses on how these components need to interact with each other. In this paper, we base our architecture on open-source and the SAS Viya ecosystem for the purpose of maintaining a reference point, but state that these points and guidelines are paramount regardless of specific software.

REQUIREMENTS FOR A PROTOCOL ANALYSIS, OPTIMISATION AND GENERATION SOLUTION

A "day in the life" of a draft protocol generation process yields the following requirements. Let us first start by outlining a broader architecture.

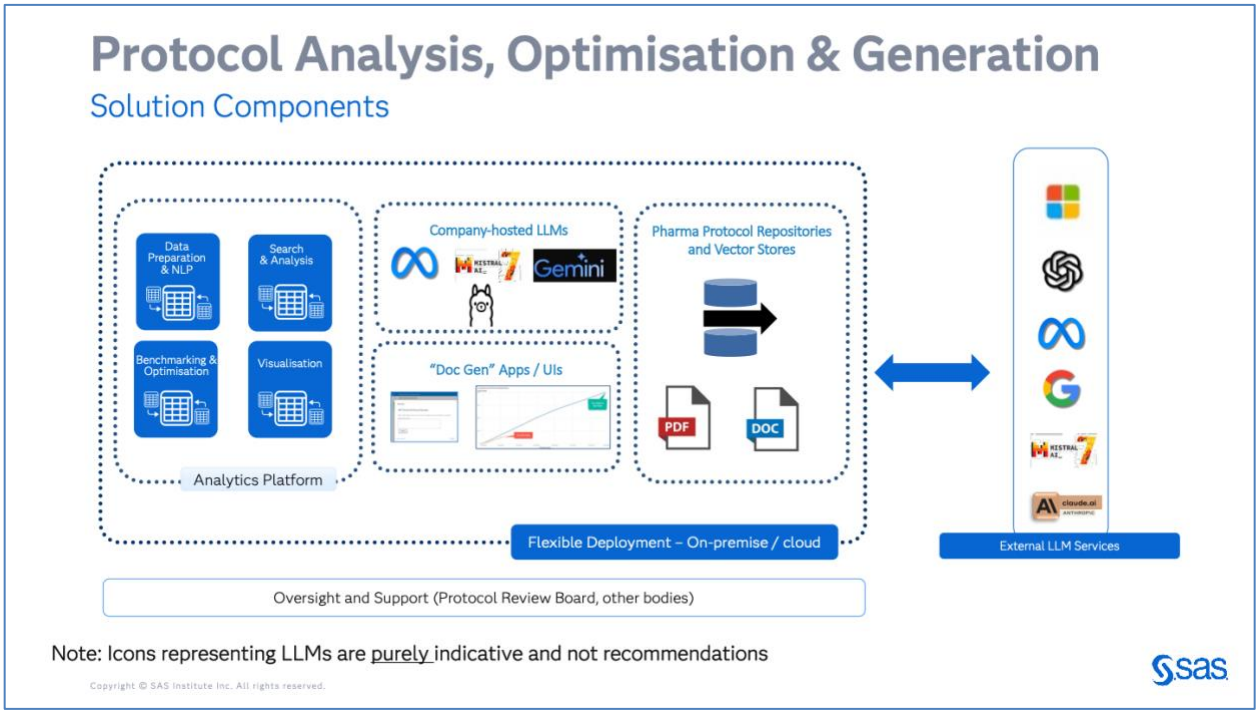

Figure 1: Solution components for a comprehensive Protocol Analysis, Optimisation and Generation solution

The broad requirements of the solution can be encapsulated as follows:

1. Look for unified solutions that cover relevant upstream and downstream tasks. The upstream tasks include data ingestion for past protocols and relevant current protocol inputs, and data structuring which make heavy use of Natural Language Processing (NLP). Downstream application should facilitate communication and feedback, such as from a protocol review board for example
2. Make considered decisions regarding Generative AI components. Protocol draft generation might require a Large Language Model, which are usually large-footprint services provided by external vendors, with some options to host them in house, or within network security boundaries that ringfence access to these models. Hosting LLMs in-house has cost and infrastructure considerations and provides some security benefits. Also, LLMs have opaque pricing which can go out of bounds if not closely monitored.

3. Consider access and auth models. Not all stakeholders require access to all components in a broad-based solution. Determine roles and privileges for various user and developer groups beforehand.

DATA INGESTION AND STRUCTURING

Clinical protocols are very lengthy documents with number of pages ranging from 50 to many 100s of pages. For a given study it is necessary to consider relevant segmentation and categorization of these artifacts using Natural Language Processing (NLP).

Natural Language Processing refers to a body of methods that involve using a combination of statistical and linguistic techniques to process unstructured data and convert it to a form that enables more accurate, relevant and focused consumption of insights. The primary techniques involved in NLP are:

i. Segmentation of text sections – this is to ensure that a long document is broken down into more manageable sections. A crude way of expressing this would be to refer to it as chunking, but chunking usually employs very simple methods focused on document length (and length of corresponding tokens). Rather, the focus of segmenting text should be on splitting out sections that carry some meaning.

ii. Noise removal and data quality – some sections of protocols may not be relevant for future studies. Also, repeated or boilerplate text only adds to token cost and increases noise for downstream analytical tasks. These should be identified through pattern analysis and removed from final structured data.

iii. Categorization: Categorization involves assigning a tag to each text content block, to classify the same as part of a logical section, the taxonomy of which is determined by domain knowledge. For example, categories in the clinical trial domain could involve eligibility criteria, endpoints, control arms and so on.

iv. Embedding: This refers to the process of converting text into a numerical representation along several dimensions. The goal of this exercise is to facilitate semantic search at downstream stages. Embedding is also called vectorization.

v. Ability to ingest various unstructured data formats – PDFs , Word documents etc.

vi. Data standardization and summarization - data across multiple protocols may follow different standards and styles. NLP methods such as text summarisation help standardise text in common language and style while retaining meaning. Large Language Models (LLMs) can also play a role in this early stage.

NLP yields analysis mechanisms which help stakeholders assess past studies in a much easier and convenient manner.

Data for clinical protocols does not consist of unstructured text alone. There are other data elements, including structured data which need to be factored into the equation. Examples include past study data comprising lab results, study outcomes, prior phase outcomes, recorded adverse events of past-generation drugs etc. You require governable and traceable Data Engineering pipelines and Data Management capabilities to seamlessly orchestrate and unify data required for a planned study.

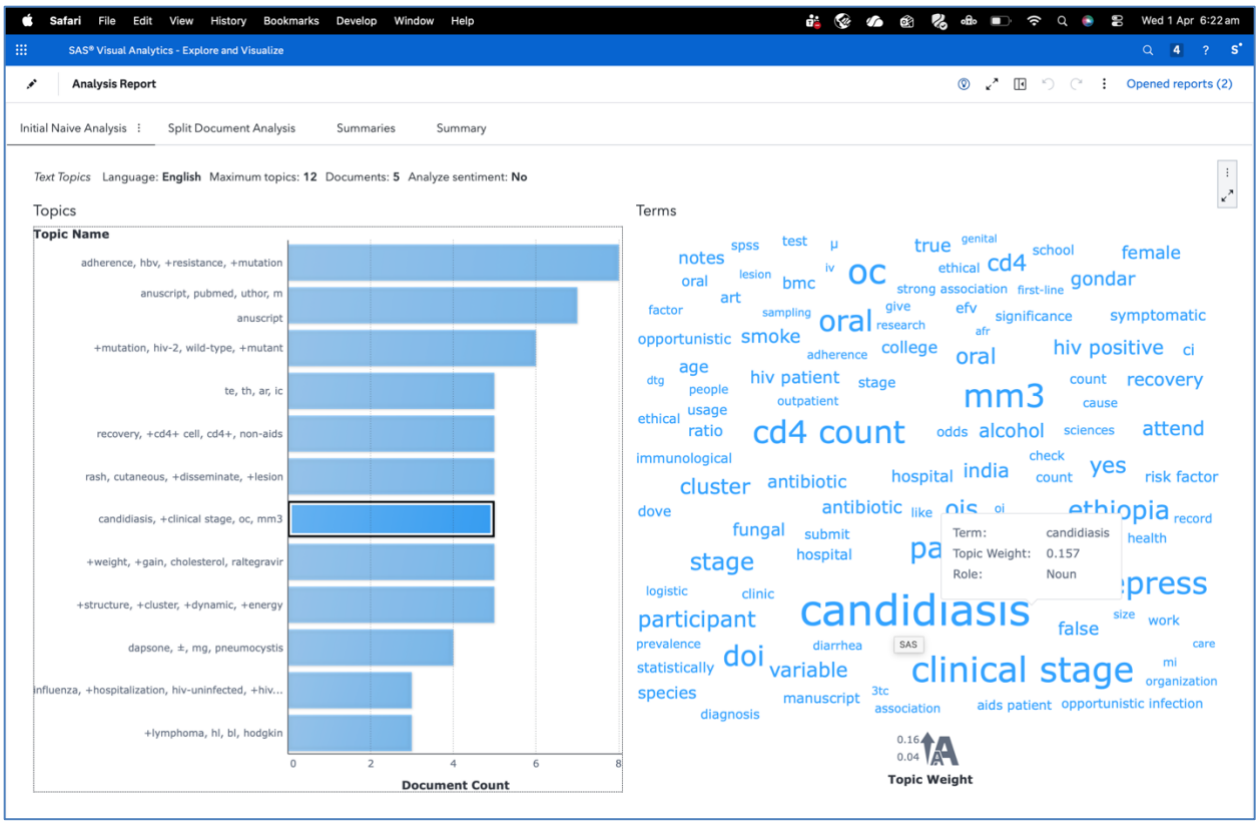

Figure 2: An example of how Topic Discovery, an NLP technique can be used to identify segments and split multiple documents based on contents

SEARCH AND ANALYSIS

A common pattern is to search a corpus of documents for past protocols in similar areas. Conventional mechanisms for this search used to rely on manual bookmarking mechanism and indices. Semantic search has made this more convenient by adding the power of vector-based search which dynamically searches for similar documents through distance-based comparisons. This allow for greater discovery of past patterns even in granular details such as for example, identifying protocols and plans which dealt with similar eligible populations even if the focus of these protocols might have been in a different area. Search mechanisms should accommodate various patterns such as vectorised (embedding distance-based) search, keyword search, indexed search etc. to offer flexibility and increase relevance.

Large Language Models (LLMs) find good applicability at this stage. LLMs are models which specialize in text generation through prediction of the next likely word or token in a sequence. In doing so, because of their training and architecture, LLMs can generate human-readable and relatable text which helps us summarise and consume the core ideas of a varied set of documents. LLMs therefore lend convenience and aid understanding in the research process.

SIMULATION, BENCHMARKING AND OPTIMISATION

The above analysis methods help you first identify a pool of likely input criteria for your target study. However, to ensure that you specify criteria that are suited to your target study, and don't just repeat past studies, you require tools that help you create scenarios and predict likely outcomes. You achieve this through simulation models that formulate study criteria by:

i. First, identify key input characteristics (which may exist in text format or LLM formatted text description)

ii. Translate these input characteristics to data elements. For this, you require intelligent mechanisms to map data from text descriptions to current data dictionaries and glossaries.

iii. Once data elements are identified, you harness either existing data sources, or synthetic data sources (based on original real-world characteristics) to model likely outcomes. Your outcomes can range from operational targets (such as likely enrolment) to final outcomes (effectiveness of intervention, primary endpoints).

iv. Having generated these predictions (often across multiple scenarios), the next step is to optimize outcomes by adjusting the variables within your control. For instance, relaxing eligibility criteria may increase enrolment. However, this can introduce confounding variables—such as healthier patient populations—that bias estimates of true treatment effectiveness (i.e., selection bias). In other cases, broader inclusion may also elevate the risk of adverse events. Optimisation helps you formalize these criteria to help you find the sweet spot you are looking for.

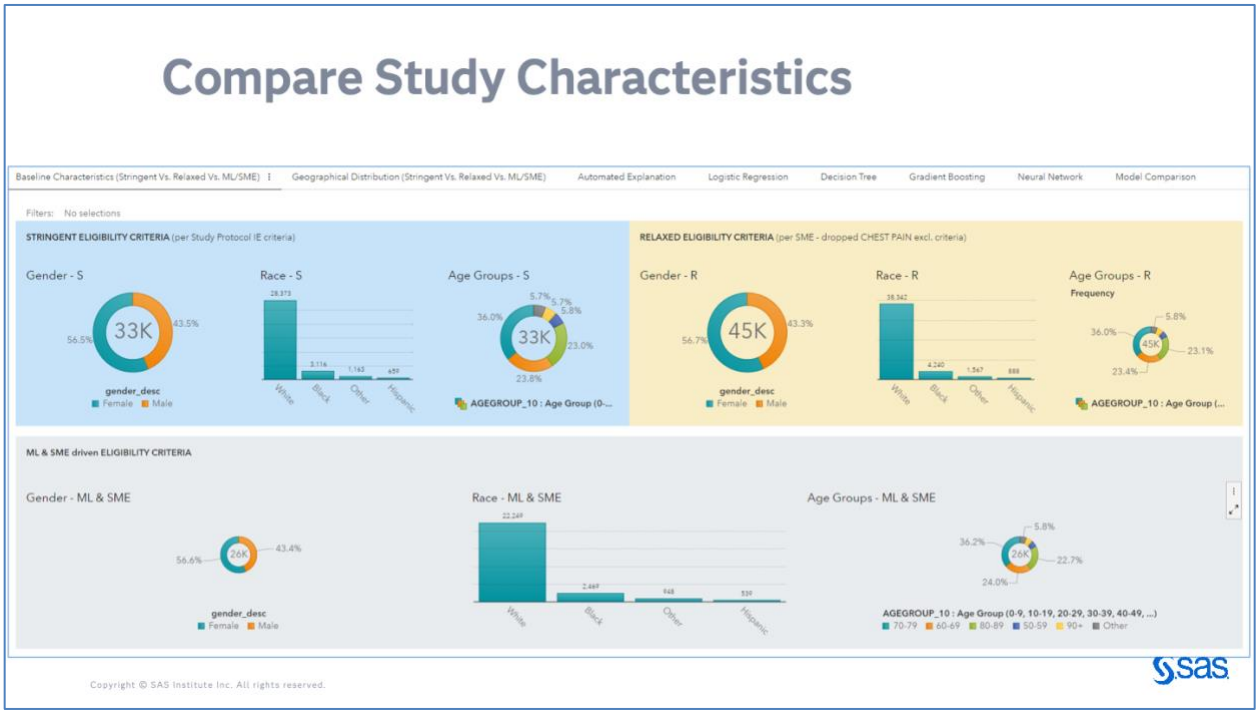

Figure 3: An example of how you can compare different scenarios to identify the right level for a control

PROJECT AND NOTEBOOK MANAGEMENT

A commonly ignored "soft skill", this involves the recording of all relevant analytical results along with supporting evidence which are eventually used for creating context. This context is then used in the downstream draft protocol generation phase.

In technology terms, the User Interface (UI) plays a huge supporting and enabling role in these situations. User Interfaces and applications help you search, analyse and visualize data conveniently, aided through automated explanations. This helps provide your project a seamless and not a disjointed flow of information and aligns with generation mechanisms down the line.

For example, modelling for outcomes after carrying out variety of predictive exercises in Model Studio, an analytical application forming part of the SAS Viya platform helps you generate a project document

capturing only the relevant part of the exercise (and, as a user, you have some flexibility in deciding what to include in the document).

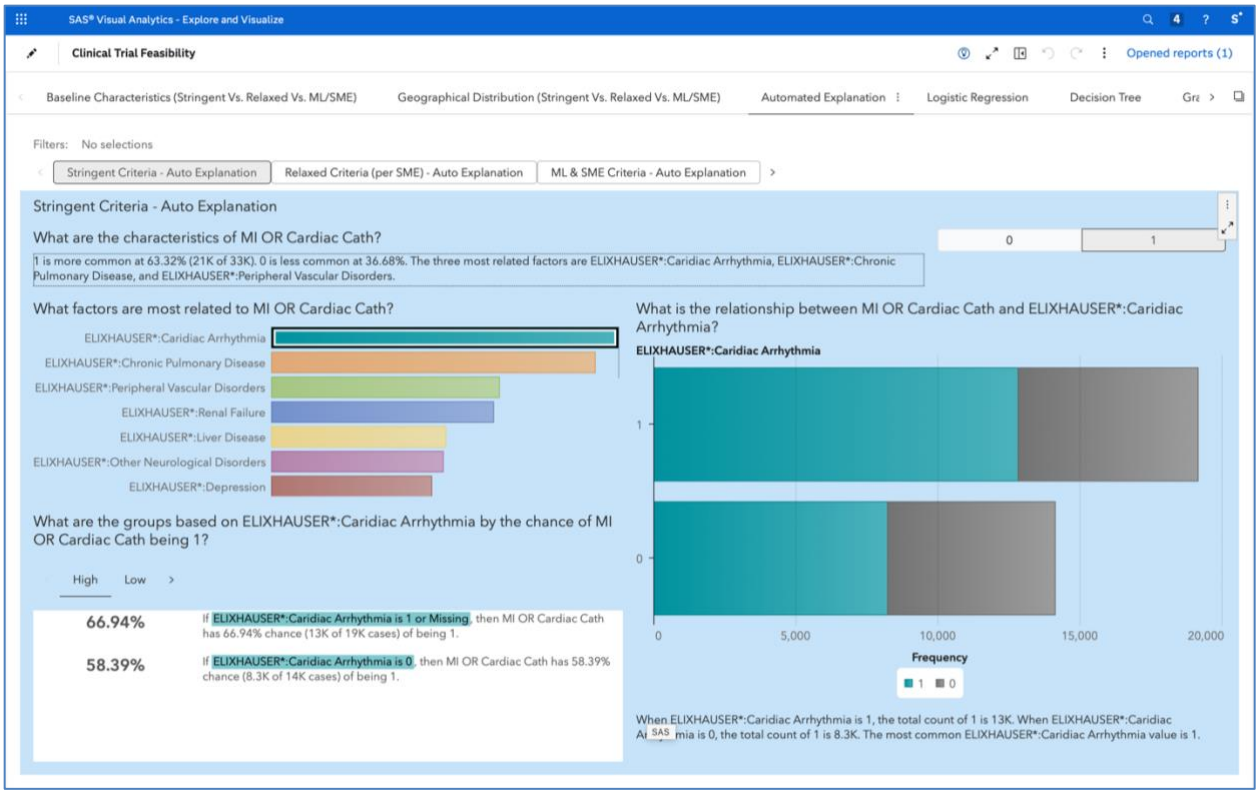

Figure 4: An example of how automated explanation can help retain insights and inputs that can be used by draft protocol generation downstream

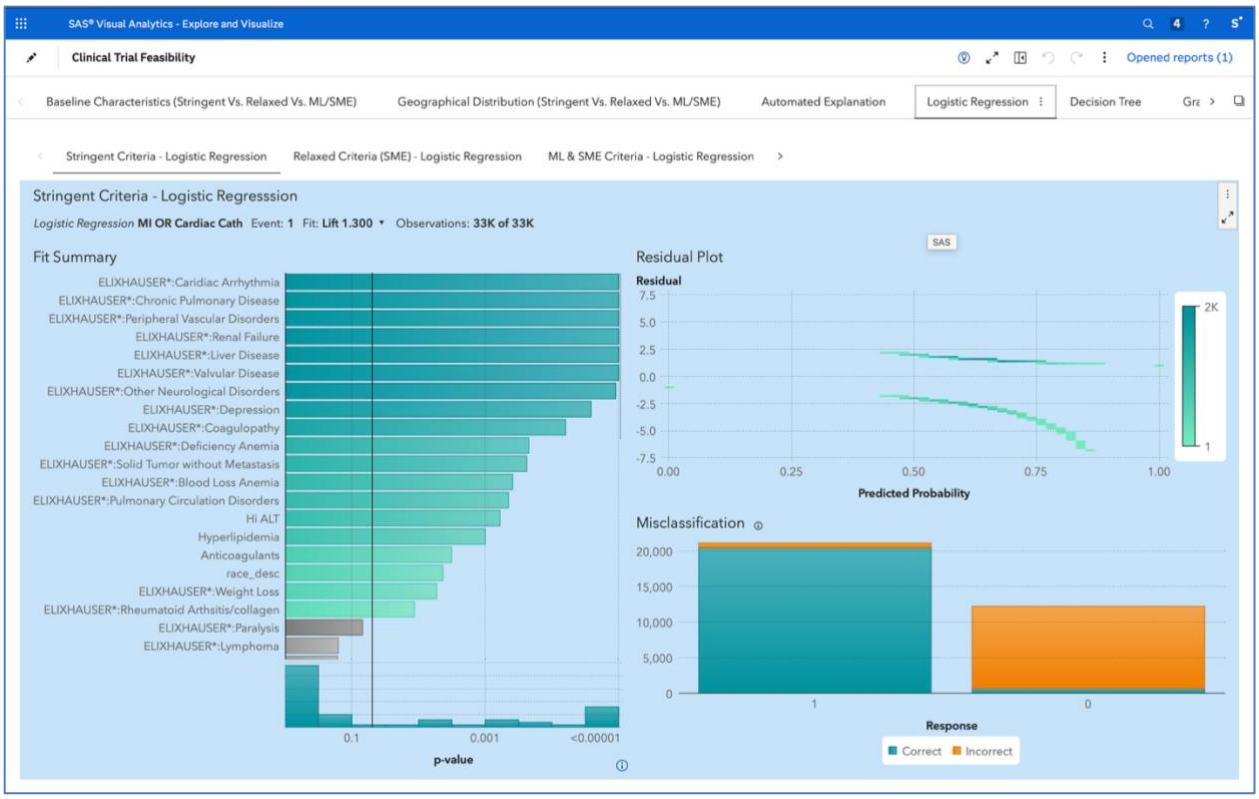

Figure 5: An example of the Analysis module

DRAFT GENERATION

The final stage is to generate the draft protocol for review. The key elements to consider at this stage are:

1. Context: Recall previous section. It's now apparent that the identification and collection of key contextual elements from your model are crucial to help generate the final output document, i.e. the protocol. Your chances of success are maximized if you were to ensure that your contextual sections align with the protocol sections that are desired. For example, the modelling results which describe the final profile (in data language) of the likely enrolees serve as the context for drafting the eligibility criteria.
2. Prompt instructions: This is the part of the generation prompt you pass on to an LLM, that clearly describes the current state and desired output. The prompt should capture the role of the generation mechanism (i.e. the LLM) the range of data available to it and granular details such as whether the LLM should attempt writing only a section of the protocol or attempt the entire protocol draft.
3. Protocol template: Protocols need to be standardized to aid review and use during the study. Your prompt should clearly articulate the template that should be used to populate the protocol (or a section therein). Templates such as the ICH M11 template can be utilised for this purpose.
4. Collaborative editing: This is crucial because of two reasons:
    a. A protocol is meant for use by multiple stakeholders whose inputs should be sought and reflected
    b. Protocol generation mechanisms, driven by LLMs, are not infallible and do tend to generate inaccurate or irrelevant content.

The correct approach would be to look at the LLM as only a tool, while humans come into the loop through collaborative editing mechanisms and refine and correct the draft protocol before finalisation. The collaborative editing mechanism should be able to handle multiple users and concurrent edits, and maintain traceability regarding who edited which section, finally leading to an approval workflow to finalise the draft for review.

PLATFORM CONSIDERATIONS

Protocol analysis solutions deal with sensitive data and require orchestration across different capabilities. They need to be able to unify data which might originate from multiple diverse source environments. Some key considerations include:

Scalability Considerations

Based on the number of protocols under development and the source data used for protocol analysis, the platform should be able to handle concurrent workstreams effectively. Protocol analysis, optimisation and generation, while regular, is not uniform in workload characteristics and volumes. Ephemeral resources which scale horizontally help ensure efficient use of resources, while also isolating data and workloads per project.

Security Considerations

Some of the data inputs that go into protocol design, especially for in-progress studies or new therapeutic areas are highly sensitive and represent intellectual property. Clear separation of roles among the different stakeholders involved should be enforced through role-based access to data, resources and technology capabilities. A major area that requires monitoring is the interaction with Large Language Models, whose inherent complexity increase the chances of data leakage and security breach. Care should be taken to ensure that LLMs are accessed using secure patterns and where sensitive data is involved, in internal environments only. Also, clear logging and reporting of activities across the platform should be enforced to identify actions and responsible parties.

User Experience and Access

LLMs promote ease of consumption but are also prone to hallucinations and inaccurate or irrelevant output. User Interfaces should retain the conversational element but also build in controls for Human-in-the-Loop aspects such as collaborative editing and approval gates prior to protocol finalisation.

Efficiency and Cost

Token usage in LLMs can easily run up costs. Prompts designed should go through a process of review and testing prior to finalisation. Where possible, use of templated prompts which are baked into applications (rather than allowing users to write their own prompts) should be enforced.

CONSIDERATIONS FOR SUCCESS

Technology components on their own are not a guarantee for success. You need to back it up with some other considerations involving design of the entire system. Some points to consider are:

PEOPLE FIRST: HUMAN-IN-THE-LOOP CONSIDERATIONS

While it's tempting to design a fully automated system from the start, a system involving AI should have gates at logical points so that humans can get into the loop and check if desired outcomes are met. When implementing your protocol analysis solution, focus on ease of use and governed solutions with mechanisms that enforce trust in results and process

COST CONSIDERATIONS

Technology plays a huge enabling role in some cases and might be overengineered in others. Balance technology against suitability for any given task and make efforts to investigate and evaluate hidden costs that might occur at various stages. A good example is that of token costs from LLM usage, which can easily go beyond thresholds if controls are not in place. Other good cost control levers which do not compromise the benefits of technology are to capitalise on common capabilities at various stages of the business flow. For example, Large Language Model usage helps not only in downstream generation but also in upstream standardisation of text data. Also pay attention on how to enable user personas through adequate and well-planned training so that they use the technology in the right way.

QUALITY FOCUS

At all points of the process ranging from data preparation to final draft generation, focus on quality by measuring and assessing the accuracy and robustness of your process. These involve setting up experiments for comparison, evaluation based on specific criteria and implementing mechanisms for early identification of retraining / model re-development needs.

CONCLUSION

Through this paper, we offer a recipe for organisations to implement an automated and efficient, data-driven protocol design and generation. Firstly, note that there is never one turnkey solution which is tuned for addressing all enterprise protocol systems. Rather, the focus should be on how well the solutions enable and adapt to the key personas and stakeholders who form the critical component of the process.

Secondly, be “curious but sceptical” of AI. AI systems provide tremendous opportunity for digital transformation and automation of process and need to be welcomed for their potential to make the protocol design process more productive. However, overengineered, poorly designed, or shortsighted AI systems work against organisation by increasing governance and monitoring costs with low effectiveness. When considering an AI transformation of an existing process, always compare with alternatives based on savings and relative cost.

Teams involved in protocol design have increasingly started adopting AI. As you evolve in your own journey, focus on how to make your technology adaptable to changes in process or output considerations that may emerge. We suggest that you will find it beneficial to consider unified technology solutions that allow you scope for customisation and interoperability with a wide range of frameworks.

REFERENCES

- Martin, Zach and Sankaran, Sundaresh, Revolutionizing Clinical Trials with Intelligent Protocol Optimization, PHUSE US Connect 2025, https://www.lexjansen.com/phuse-us/2025/ml/PAP_ML08.pdf
- First-pass yield: quick overview, Wikipedia, https://en.wikipedia.org/wiki/First-pass_yield
- Natural Language Processing (NLP) Outline, SAS Institute, https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html
- ICH M11 Protocol Template Guidance, https://www.ema.europa.eu/en/ich-m11-guideline-clinical-study-protocol-template-technical-specifications-scientific-guideline

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sundaresh Sankaran
SAS Institute
Phone : +1 919 400 3266
Sundaresh.sankaran@sas.com
https://www.linkedin.com/in/sundareshsankaran/

Sherrine Eid
SAS Institute
Phone: +1 919 400 3266
Email: Sherrine.Eid@sas.com
https://www.linkedin.com/in/sherrineeid/