Paper TA02

Future-Proofing Clinical Data Sciences: The Pivotal Role of Therapeutic Area Proficiency

Sascha Ahrweiler, PHUSE LLC/Bayer AG, Wuppertal, Germany

ABSTRACT

Understanding Therapeutic Areas (TA) is becoming increasingly critical for clinical data scientists in the pharmaceutical industry, particularly as artificial intelligence continues to revolutionize the field. This paper introduces an innovative AI-driven methodology, utilizing ChatGPT technology, for generating educational articles on TAs. These articles are intended for publication on the PHUSE Education website, serving as a peer-reviewed, continuously updated knowledge repository. Emphasizing the role of TA expertise as a career-defining skill in the next 5-10 years, the paper underscores the impending transition from manual programming to specialized data interpretation and analytics. The initiative aims to foster self-learning and knowledge dissemination, aligning with the PHUSE Education mission of creating proficient practitioners capable of teaching others, thus enhancing the data science community's competence in this critical domain.

INTRODUCTION – MY FIRST STEPS IN CLINICAL DATA SCIENCE

When I first started in clinical data science, my journey began with a job as a SAS® programmer during my student days. In my math studies at university, I had a course on SAS V6.12 that lasted six weeks. This course was a real eye-opener for me. I already had a good handle on programming, but learning SAS®, with its focus on statistics, was a completely new and thrilling experience.

Soon after I finished this course, I came across a job opening at the University's Institute for Statistics in Medicine. They were looking for a student assistant with SAS® skills. I applied and, to my surprise, I got the job. My initial task there was related to a research project on dental implants, a large-scale study that spanned over 10 years and gathered detailed data, including the placement of dental implants in the mouth.

My responsibility was to create boxplots that illustrated the marginal jawbone level changes measured in distal and mesial over time. I put together these charts and they looked impressive, but my supervisor quickly pointed out a crucial issue. Despite their visual appeal, they didn’t accurately represent the data. I had mixed up some data points and overlooked key information, such as missing data and inaccuracies. It also turned out that I had no clue what mesial and distal meant to properly check the accuracy of my plots on my own.

This experience taught me an invaluable lesson that has guided me through my 25-year career in clinical data science. Beyond just programming skills, it's essential to deeply understand the data you are working with. Moreover, this experience highlighted the importance of having knowledge in the specific therapeutic area, like in my case, dentistry. This knowledge is crucial for a clinical data scientist because it ensures that the data is not only accurately represented but also meaningful in the context of the specific field. This understanding of the therapeutic area, combined with programming skills, is what truly makes effective and impactful clinical data science.

WHY THERAPEUTIC AREA KNOWLEDGE IS MORE IMPORTANT THAN EVER

In the ever-evolving world of clinical data science, the importance of specialized knowledge in therapeutic areas is becoming increasingly crucial. This shift is particularly evident in the wake of transformations the field has undergone recently. Even before the COVID-19 pandemic, challenges were emerging due to the trends of offshoring or outsourcing programming activities. Post-COVID, these challenges have been accelerated by financial constraints leading many companies to implement additional cost-cutting measures.

Additionally, following the current hype of large language models, the landscape of clinical data science is being reshaped by digital disruption, with technologies like generative AI beginning to replace entry-level programming roles. This technological advancement, along with the increasing use of new data sources like device data in clinical trials, is changing the face of the field.

Amidst these changes, clinical data scientists are faced with the question of how they can increase their value to their companies. A key part of the answer lies in the combination of an intimate understanding of data in specific therapeutic areas with the technical skills to analyze and present this data effectively. This blend of domain-specific knowledge and technical prowess is a differentiating skill that can set a clinical data scientist apart.

Educational resources are vital for acquiring this specialized knowledge. An excellent example is an educational article written by Rabia Ahmed and Alberto Montironi for PHUSE Education. Their article offers in-depth insights into what clinical data scientists should know about Oncology, demonstrating the value of such knowledge. The article, a valuable resource for anyone in the field, is accessible here.

Having a foundational medical understanding in a specific therapeutic area, like oncology, enables clinical data scientists to be less dependent on medical and statistical guidance. It allows for more informed conversations with medical professionals and statisticians and prepares them for the intricacies of regulatory submissions. Moreover, it equips them to recognize and navigate the potential pitfalls in their data, ensuring they are aware of the unique challenges each set of data presents.

As clinical data science continues to adapt to new challenges and technologies, the value of having in-depth knowledge in a therapeutic area cannot be overstated. It's about merging the art of programming with the science of the data's context, making clinical data scientists not just coders, but informed and valuable contributors in their field. As this paper will demonstrate, the implicit human knowledge based on experience currently still outrivals other explicit knowledge building such as artificial intelligence, and could become a distinguishing value generation factor for clinical data scientists.

HOW ARTIFICIAL INTELLIGENCE CAN HELP YOU TO ACQUIRE THE THERAPEUTIC AREA KNOWLEDGE YOU NEED

The question arises though: how can a clinical data scientist acquire knowledge in a specific therapeutic area (TA) without undergoing extensive medical training? Interestingly, the same generative AI that poses a threat to certain aspects of clinical data science could also be a boom in this context.

Generative AI, often perceived as a threat due to its potential to automate entry-level programming tasks, can also be a powerful ally in enhancing productivity and knowledge acquisition. This dual nature of AI highlights its potential as a tool to bridge critical knowledge gaps in therapeutic areas for clinical data scientists.

The key lies in using AI to acquire TA knowledge quickly and efficiently. By effectively utilizing AI, clinical data scientists can close a crucial gap in their understanding, making them more versatile and knowledgeable in their field. This approach involves understanding the do's and don'ts of prompting AI tools to extract relevant and accurate information.

METHOD AND STRUCTURE OVERVIEW

I will follow the structural approach following the structure provided by the PHUSE Educations Therapeutic Area cluster. This offers a comprehensive framework for understanding various aspects of a therapeutic area. The PHUSE framework can also be accessed here.

For practical purposes and based on my own TA experience, I will focus on considerations for good and bad prompts using a specific use case of Parkinson's Disease. The complete written educational article can be found on the PHUSE Education Website in the Therapeutic Area cluster (PHUSE TA PD). It is important to understand though, that this output is the end outcome of a series of peer reviews. The initial draft has been created by using similar prompts as described further below.

Using this structural approach on Parkinson's Disease, the following areas should be covered to acquire appropriate TA expertise related to clinical data science activities:

- **Description of Disease:** A comprehensive overview of Parkinson's Disease (PD).
- **Demographics and Baseline Characteristics:** Insight into the patient demographics and baseline characteristics specific to PD.
- **CDISC Standards and Therapeutic Area User Guides:** Guidelines and standards relevant to PD, as per CDISC.
- **Agency Guidelines:** Regulatory guidelines and frameworks specific to PD.
- **Study Design and Study Endpoints:** Common approaches and endpoints used in PD research.
- **Data Challenges:** Identification of typical data challenges encountered in PD studies.

In the following, this paper will explore how ChatGPT 4 (www.openai.com) with a Plus subscription can be utilized to create outputs for each of these structured areas, providing a comprehensive guide on harnessing AI for enhanced understanding and analysis.

GETTING STARTED – GENERAL PROMPT ENGINEERING AND SYSTEM INITIALIZATION

Utilizing a generative AI tool such as ChatGPT requires the user to precisely instruct the AI to create a desired result. In this context, stating precise instructions will be referred to as prompting or prompt engineering. Before starting our journey to utilize generative AI to acquire TA knowledge, understanding the do's and don'ts of prompting AI tools to extract relevant and accurate information is crucial.

DO'S FOR PROMPTING GENERATIVE AI TOOLS:

1. Be Specific: Clearly define the therapeutic area and the specific information you need. For example, if you're working on oncology, ask for insights related to cancer types, treatments, and recent research findings in that area.
2. Ask for Sources: Request the AI to provide sources or suggest readings where you can find in-depth information.
3. Update Requests: Keep your queries current by asking for the latest research and developments in the therapeutic area.
4. Check for Relevance: Ensure that the information provided by AI is relevant to your specific clinical data science context.

DON'T'S FOR PROMPTING GENERATIVE AI TOOLS:

1. Avoid Vague Requests: Generic questions will lead to generic answers, which might not be useful.
2. Don't Overlook Verification: Always cross-check the information provided by AI with reliable sources.
3. Don't Rely Solely on AI: Use AI as a supplementary tool, not as the sole source of your therapeutic area knowledge.
4. Avoid Misinterpretation: Be cautious of how you interpret the data and information provided by AI; misunderstandings can lead to incorrect applications in your work.

I will be more specific about the implementation of these do's and don'ts and will provide examples of good and bad prompts in each section of this paper. Before diving directly into the actual creation of the educational article for a TA though, it is beneficial to start from a clean chat and provide the overall context to ChatGPT for the creation of the article. This can be achieved by either directly referring to the PHUSE Education article on Oncology or providing the details in general settings.

A prompt, which would initialize a ChatGPT session could look as follows.

EXAMPLE OF AN INITIALIZATION PROMPT

"I am a clinical data scientist in a pharmaceutical company. I want to write an educational article about Parkinson's Disease to acquire relevant knowledge about this therapeutic area. My article should use the following structure:

- *Description of Disease: A comprehensive overview of Parkinson's Disease*
- *Demographics and Baseline Characteristics: Insight into the patient demographics and baseline characteristics specific to PD.*
- *CDISC Standards and Therapeutic Area User Guides: Guidelines and standards relevant to PD, as per CDISC.*
- *Agency Guidelines: Regulatory guidelines and frameworks specific to PD.*
- *Study Design and Study Endpoints: Common approaches and endpoints used in PD research.*
- *Data Challenges: Identification of typical data challenges encountered in PD studies.*

You will help me to draft this article. The language should be easy to understand for a clinical data scientist and not utilize medical terms if that is not necessary. Always provide reliable sources for your output so that the interested reader can do further research.

Please let me know if you understand this task and ask clarifying questions to better understand the context."

By using this prompt, ChatGPT receives all the required context and instructions to provide fairly reasonable output. With the current ChatGPT version, this prompt could either be stated directly in a new chat or alternatively could be added to custom instructions. The latter will make this instruction available for other use cases as well.

DESCRIPTION OF DISEASE

In this first TA specific section, we'll explore how to use generative AI to learn about Parkinson's Disease (PD) effectively. To illustrate this, I will start to examine the difference between a poorly crafted prompt and an effective one. I will then use the effective one to further refine it to an even more effective prompt.

Example of a Poor Prompt
"*Tell me about Parkinson's Disease."*

This is a bad prompt because it is too general and doesn't provide any specific context. It would likely result in a broad, non-specific overview of PD that lacks depth and relevance for a clinical data scientist. This kind of prompt does not guide the AI to provide information that is pertinent to a clinical data scientist's need.

Why This Prompt Is Ineffective:
- It does not specify the need for information relevant to clinical data science.
- It lacks a request for epidemiological data and progression details, crucial for PD research.
- It does not ask for links to trusted sources for further, in-depth reading.

Crafting an Effective Prompt:
A better prompt would be: *"Provide a detailed overview of Parkinson's Disease, focusing on aspects relevant to clinical data science. Include information on epidemiology with high-level statistics on affected populations, typical progression of the disease, and common data types used in clinical trials for Parkinson's research. Also, suggest links to trustworthy sources for in-depth understanding, suitable for someone with a basic medical background."*

Why This Prompt Is Effective:
- It sets a clear context, asking for information tailored to the needs of a clinical data scientist with limited medical background.
- It requests specific information on epidemiology, disease progression, and data types in clinical trials, which are directly relevant to a clinical data scientist's role in PD research.
- It asks for high-level explanations and links to authoritative sources, ensuring the information is accessible yet comprehensive.

Utilizing the Good Prompt: Using this well-structured prompt, we can gain a comprehensive understanding of PD relevant to clinical data science. This includes:

- Epidemiological data such as prevalence, incidence, and affected demographics.
- An overview of the disease's progression, including key stages and symptoms.
- Insights into the types of data commonly used in PD clinical trials, such as patient-reported outcomes, clinical scores, and imaging data.
- Trustworthy sources for further reading, including patient communities, official guidelines, and standards organizations.

Crafting an optimal prompt
While the last prompt is already much better than the bad prompt, we can even utilize ChatGPT to further optimize the much better prompt by asking ChatGPT the following question: *"Is there anything I can do to make the good prompt even better? Ask me 5 more questions to further clarify the context. Once you are completely sure about the context, provide the perfect prompt to reach our goal"*

This approach and starting a dialogue with ChatGPT to clarify the context will lead to a refined prompt, which could look like
"Provide a concise overview of Parkinson's Disease, focusing on aspects most relevant to clinical data science. Highlight existing treatment options and current research areas in PD. Summarize the current challenges faced in PD research. Include a variety of links for deeper exploration, encompassing scientific articles, patient communities, official guidelines, and multimedia resources like informative YouTube videos. The information should be high-level, suitable for a clinical data scientist with limited medical background, and should serve as a springboard for more detailed study."

Why this prompt is even more effective:
- **Targeted Information:** It specifically asks for current treatment options and research areas, aligning with your interest in understanding the latest developments in PD.
- **Appropriate Depth:** The prompt requests a brief overview but emphasizes the need for links to detailed resources, allowing for deeper exploration at your convenience.

- **Relevance to Clinical data science**: It maintains a focus on the aspects relevant to your field without delving into unnecessary medical complexities.
- **Inclusion of Challenges:** By asking for current challenges in PD research, it ensures you stay informed about the latest issues and areas of focus in the field.
- **Resource Diversity:** The prompt calls for a range of resources, from scientific articles to YouTube videos, catering to different learning preferences and providing a well-rounded understanding.

The above-mentioned prompt examples should lead as general guidance, on how to effectively work with a generative AI tool such as ChatGPT to create a reasonable educational output for a clinical data scientist to learn more about the fundamentals of a specific disease. The actual output for the specific PD educational article for this section can be found (PHUSE TA PD). As we will see in the following, the deeper we will dive into clinical data science challenges, the more challenging the prompt engineering will be.

DEMOGRAPHIC AND BASELINE CHARACTERISTICS

After we have successfully learned the principles of how to create an effective prompt to learn about the foundations of Parkinson's Disease, the section about demographic and baseline characteristics might be a little bit more challenging. The goal of this educational article is to understand which Demographic and Baseline characteristics are usually measured in a clinical trial for PD research. This requires even some more context about clinical trials.

In this specific section, we want to understand, how a population in a clinical trial for Parkinson's Disease would be defined. Besides the general demographic variables, such as age, gender, height, weight, etc. we want to learn how a patient suffering from PD could be described. For example, we want to understand if there are specific measurements, which describe the state of the disease. This understanding will also be very important for the later section of clinical study design and endpoints. Therefore, it is of high importance to do this right straight from the beginning.

Example of a poor prompt for the Characteristics in Parkinson's Disease
"What are the Demographic and Baseline Characteristics for Parkinson's Disease?"

Why This Prompt Is Ineffective:
- **Lacks Specificity:** This prompt is too vague. It doesn't specify what kind of information is needed.
- **No Focus on Relevance to Clinical Data Science and clinical trials:** The prompt does not guide the AI to provide information that would be useful specifically for statistical analysis or research.
- **Misses Key Aspects:** It doesn't mention any interest in variables, which are usually acquired in clinical trials.

Creating an Effective Prompt
To develop a good prompt, consider the following methods:

1. **Be Specific about Data Types:** Clearly state the type of demographic and baseline characteristics data you need. For example, include age, gender, ethnicity, geographic distribution, or socio-economic status, medical history and endpoints usually measured in a clinical trial on PD.

2. **Mention Relevance to Statistical Analysis:** Indicate that you need this information for clinical data science purposes to understand the clinical trials and the respective data. This guides the AI to tailor the information to be more relevant for data analysis.

3. **Ask for Baseline Characteristics:** Specify that you're interested in baseline characteristics of PD patients, such as common comorbidities, general health status before PD onset, or typical initial symptoms.

4. **Request Links to Authoritative Sources:** Ask for links to research papers, surveys, or databases that provide detailed demographic data and baseline characteristics of PD patients.

Example of an Effective Prompt
"Provide a detailed analysis of the demographic and baseline characteristics of Parkinson's Disease patients, which are usually measured in clinical trials. Focus on aspects relevant to clinical data science. Include information on variables, which are acquired in a clinical trial to describe the study population. Also, describe typical baseline characteristics to describe the study population as detailed as possible. This should include the primary and secondary variables measured at study screening or study start. Suggest authoritative sources for in-depth demographic data and baseline characteristics suitable for statistical analysis in Parkinson's Disease research."

This prompt should elicit a response that is directly relevant to a clinical data scientist's need, providing a detailed and focused overview of the demographics and baseline characteristics of PD patients, along with reliable sources for further research. The actual output for the specific PD educational article for this section can be found (PHUSE TA PD). As you can see, the created output appears to be reasonable and mentions the important endpoints such as the Unified Parkinson's Disease Rating Scale (UPDRS) for example. Nevertheless, a proper peer review should be conducted to confirm the correctness of the provided information.

CDISC STANDARDS AND THERAPEUTIC AREA USER GUIDES

In this next section, we will now further increase the clinical data science relevant level of details and will try to learn more about the CDISC data standards specific for PD. The adherence to industry-relevant data standards, such as the CDISC standards for SDTM and ADaM is crucial for a Clinical Data Scientist. This is also true in our example of Parkinson's Disease (PD) research. A well written educational article should provide insights into existing general CDISC data standards and should state if a respective Therapeutic Area User Guides (TAUGs) exist. However, accessing and utilizing these standards can be challenging for a tool such as ChatGPT, since this information are not freely accessible. Utilizing AI to increase this specific TA expertise requires therefore some additional effort.

Firstly, it's important to note that CDISC Standards are protected behind a firewall and are accessible only to CDISC members. For the creation of this educational article, we assume though that the user is a CDISC member for example through their corporate membership and with this has access to the standard library. In either case, these standards are integral to ensuring consistency, quality, and compliance in clinical data regardless of the respective therapeutic area.

Once membership and website access are secured, the next step involves navigating the CDISC website to locate the general CDISC standards (CDISC SDTM, CDISC ADaM). This is a crucial phase as it lays the groundwork for understanding the overarching framework within which specific therapeutic areas, including PD, are addressed. After locating the general standards, researchers should focus on identifying and accessing the TAUGs relevant to PD. A general overview of all TAUGs can be found here.
As one can see based on the overview, there is a PD specific User Guide, which can be accessed here.
These guides provide detailed instructions and standards specific to Parkinson's Disease, making them an invaluable resource for researchers.

While ChatGPT cannot access these standards easily, utilizing such AI tools to gain further insights from these documents can be highly beneficial. For example, after downloading the relevant TAUGs as PDFs, they can be uploaded to ChatGPT, which can then be queried to extract and synthesize information. This approach can significantly streamline the process of digesting these comprehensive and often complex documents. It is also really helpful to create high level summary information for other users as a kind of springboard to acquire more in depth knowledge of these critically important documents.

Again, there are pitfalls to be aware of. General document queries, like "*Are there specific considerations for the Study Data Tabulation Model (SDTM) for PD?*" can result in incomplete or non-specific responses. To mitigate this, it is crucial to approach AI queries again with precision and provide information, which is not accessible to ChatGPT. Clearly explaining the content of the document and pinpointing where in the document specific SDTM considerations for PD can be found will yield more targeted and useful results. This specificity not only aids in better understanding but also in the practical application of these standards in PD research.

The following example again helps to understand how to build effective prompts to achieve your goal.

Example of a Bad Prompt
"Tell me about CDISC standards for Parkinson's Disease."

Why This Prompt Is Ineffective:
- It's too vague and doesn't specify what aspects of the CDISC standards are needed. The user should ask more specifically for SDTM or ADaM standards.
- ChatGPT does not have access to the requested information due to the CDISC membership restrictions.

Methods to Create a Good Prompt
Creating a better prompt requires more steps than in the earlier sections. As mentioned, the CDISC standards and the therapeutic area user guides require that you have access as a CDISC member to the CDISC website. If that is ensured, you can follow these steps:
1. Download the CDISC standard documents for SDTM, ADaM and the PD specific TAUG

2. Upload the downloaded PDFs to ChatGPT: provide the high-level content of the uploaded documents.
3. Specify the Focus on Parkinson's Disease: Clearly state that you're seeking information on SDTM or ADaM CDISC standards specifically as they relate to PD.
4. Ask for Details on Clinical Trials: Indicate that you need details on how these standards are applied in PD clinical trials.
5. Request Examples: Ask for specific examples or case studies that illustrate the application of these standards in PD research.
6. Seek Specific Sections of the Documents: Direct the AI to focus on specific sections of the uploaded documents that are most relevant to PD.

Example of a Good Prompt
"Briefly explain what CDISC Standards and Therapeutic Area User Guides (TAUGs) are. Provide links to the CDISC website where the standards and TAUGs can be accessed. Discuss the challenge of CDISC Standards being behind a firewall and accessible only to members. The attached documents contain information about Parkinson's Disease specific therapeutics are user guides, focusing on their application in clinical trials for clinical data science. Highlight key sections of the uploaded documents that are particularly relevant to PD research. Explain how these standards guide data management and analysis in PD clinical trials, and provide examples or case studies if available. The PD specific SDTM domains are described in Section 2. Provide a bullet point list for each domain and add a high-level description about the importance and the content of this SDTM domain."

This prompt is designed to yield a focused and informative response that directly addresses the needs of someone working on clinical data science in PD clinical trials, providing clear guidance on how to navigate and utilize CDISC standards in this context. As the above example demonstrates, the description of the CDISC standards and especially the TAUG requires some additional steps and very detailed knowledge for effective prompting.

AGENCY GUIDELINES

In the realm of clinical data science, particularly when dealing with diseases like Parkinson's Disease, it's essential to have a comprehensive understanding of relevant agency guidelines. These guidelines often extend beyond the specific disease to encompass the broader therapeutic area, in this case, neurodegenerative disorders. As a consequence, it might occur that some agencies do not have published a specific guidance for Parkinson’s Disease and the guidance for neurogenerative disorder should be followed. For companies, who strive for global agency submissions, the situation is even more complex, since some agencies might have specific PD guidances, while others do not.

This understanding is crucial for the design and execution of clinical trials, ensuring compliance and efficacy in research and therefore is again very useful knowledge for a clinical data scientist, who wants to be prepared for global submissions. In contrast to the above mentioned CDISC standards, all agency guidelines should be freely accessible. When trying to work with generative AI, it is still important to define the context of the information we are looking for. This again requires some initial understanding of the specific disease and the regulatory environment.

The Scope of Agency Guidelines: While specific guidelines for PD might not always be available, guidelines for neurodegenerative disorders often provide valuable insights. These overarching guidelines can include recommendations on clinical trial design, data management practices, and safety reporting standards. When utilizing AI to increase TA expertise, it's important for Clinical Data Scientists to recognize that disease-specific guidelines might be embedded within these broader directives and should be included in the AI prompts.

Global Variations and Their Impact: The landscape of agency guidelines is not uniform across the globe. For instance, the US Food and Drug Administration (FDA) might have different requirements compared to the European Medicines Agency (EMA), the Japanese Pharmaceuticals and Medical Devices Agency (PMDA), or health authorities in China. These variations can affect many aspects of clinical trials, from patient recruitment criteria to data reporting standards. Understanding these differences is vital for international clinical studies and for ensuring global compliance.

The Importance of Context and Prior Knowledge: To navigate these guidelines effectively with the support of AI, a-priori knowledge and a detailed understanding of the specific disease context are essential. A Clinical Data Scientist, who wants to utilize AI, should provide the right scope for the agency guidelines and also include the variations and their impact to receive an appropriate response. Examples of bad and good prompts are displayed in the summary table.

These considerations are helpful in creating good and bad examples for prompt engineering.

Example of a Bad Prompt

"Tell me about the agency guidelines for Parkinson's Disease."

This prompt is way too vague and does not specify the need for understanding guidelines within the broader context of neurodegenerative disorders or across different global agencies.

Example of a Good Prompt
"Provide an overview of global agency guidelines relevant to neurodegenerative disorders, with a specific focus on Parkinson's Disease. For each agency, extract the three most important points to consider for a clinical data scientist. Discuss if there are any PD-specific guidelines or those applicable to neurodegenerative disorders as a whole from agencies like the US FDA, European EMA, Japanese PMDA, and Chinese health authorities. Explain how these guidelines might vary and their implications for clinical data management and trial design in PD research."

This good prompt again can be further refined:

- **Specify Agencies:** Clarify that you are interested in guidelines from specific regulatory agencies and their potential differences.
- **Ask for Contextual Application**: Request information on how these guidelines apply to clinical trials and data management in PD within the broader scope of neurodegenerative disorders.
- **Include Examples:** If available, ask for examples or case studies where these guidelines have been applied in PD research.
- **Instruct ChatGPT to further specify context:** as demonstrated in the first section, this method might help to create an optimal prompt.

By following this structure and using an effective prompt, you can draft a comprehensive section on agency guidelines that will be insightful for clinical data scientists working in the field of PD and neurodegenerative disorders. This approach ensures a nuanced understanding of the regulatory landscape and its impact on research and clinical trials in this therapeutic area.

STUDY DESIGN AND STUDY ENDPOINTS

This section of the paper now delves very deep into the specifics of study designs and endpoints PD research. Designing clinical trials and selecting appropriate endpoints to measure the safety and efficacy of a new treatment is a very complex task. This task requires thorough expertise in drug development and cross-functional expertise from experienced statisticians, medical teams, etc. Consequently, preparing an educational article for this section requires a lot of a-a-prioriprior knowledge to effectively utilize generative AI for drafting.

The very least the writer of this section should do is to completely understand the output of the section about "Demographic and Baseline characteristics". Especially the section on baseline characteristics should provide insights into the main instruments used in clinical trials for Parkinson’s Disease, such as the Unified Parkinson’s Disease Rating Scale (UPDRS) or the Parkinson’s Disease Questionnaire (PDQ-39). Understanding these instruments and their applications is key to identifying appropriate endpoints for measuring treatment effectiveness during the conduct of a study.

Furthermore, the user should have a basic understanding of general study designs in clinical trials. Different study designs, such as randomized controlled trials, cohort studies using observation data, or case-control studies, can be utilized to demonstrate the effectiveness of new treatments. The choice of design depends on factors like the stage of the disease, expected outcomes, and patient characteristics.

These described challenges demonstrate nicely, that using AI to establish TA knowledge has its clear limitations. While AI can generate information, its effectiveness is limited by the specificity of the prompts and the a-priori knowledge of the user. A well-informed prompt that draws on the outcomes of previous sections such as the demographics and baseline characteristics will yield more relevant and accurate information. Given the complexity of clinical research, additional mentorship and guidance from more experienced clinical data scientists or their managers can be invaluable in crafting effective prompts and understanding study designs and endpoints.

Let’s try to again provide examples of good and bad prompts.

Example of a poor prompt
"Tell me about the study design and study endpoints of PD trials"

This prompt is too general and does not specify the need for understanding how different designs relate to measuring treatment effectiveness based on baseline characteristics.

Example of a better prompt
"Provide an analysis of different study designs used in PD research, focusing on how these designs measure the progress of baseline characteristics identified in earlier sections. Highlight study endpoints that are commonly used to assess the effectiveness of new treatments in PD, and discuss their relevance to the disease's progression and patient outcomes."

This prompt is more specific, asking for an analysis of study designs in the context of measuring treatment effectiveness against established baseline characteristics. It directly ties the request to previously acquired knowledge, making it more likely to yield detailed and relevant information. However, it is still too generic to receive an appropriate response.

Example of a really good prompt:
"Provide a detailed exploration of various study designs commonly employed in PD research, including randomized controlled trials and observational studies. Discuss the rationale and implications of each design, with an emphasis on control groups, blinding, and study duration.

Delve into the selection and significance of study endpoints in PD trials. Cover both primary endpoints, like motor symptom evaluation through scales like the Unified Parkinson's Disease Rating Scale, and secondary endpoints, such as non-motor symptom assessment and quality of life measurements. Address the challenges in measuring, interpreting, and statistically analyzing these endpoints.

Highlight the emerging role of digital biomarkers in PD research. Discuss the integration of wearable technologies and mobile applications for continuous symptom monitoring, their impact on study designs, and the statistical considerations for data analysis.

Examine the use of neuroimaging techniques (MRI, PET scans) and biomarkers in understanding PD pathology, progression, and treatment response. Explore the statistical methods and challenges in analyzing neuroimaging and biomarker data.

Conclude by discussing the future directions in PD clinical trial design, focusing on patient-centered approaches, adaptive trial designs, and the incorporation of real-world data."

This longer prompt includes a lot of knowledge and is an outcome of in-depth familiarization with the respective TA. This prompt might be an outcome of a series of initial prompts using above mentioned methods such as context assessment by ChatGPT. Following the motto "the journey is the reward", acquiring such specific knowledge via a dialogue with AI to craft an effective prompt is already a great outcome of a learning journey and should not be underestimated.

In conclusion, this section emphasizes and demonstrates the limitations of using AI for this use case. It also highlights the importance of detailed, context-specific knowledge in PD research. It serves as an example of how effective prompting, combined with a thorough understanding of the subject matter, can overcome the limitations of generative AI in producing meaningful and useful content.

DATA CHALLENGES

While the previous section already demonstrated the limitations of using AI to acquire TA knowledge, crafting an educational article about data challenges in PD clinical trials is an even more complex task, particularly for those who haven't engaged in hands-on clinical data science in such studies. This task can be likened to writing a book about riding a bicycle without ever having ridden one; it's a largely theoretical exercise that may not fully capture the practical nuances. Nevertheless, with the right approach and guidance, it is possible to construct a meaningful and informative piece.

The process of comprehending the data challenges in PD clinical trials primarily from a theoretical standpoint involves a deep dive into several interrelated aspects:

- **Complexity of PD Data**: PD trials generate complex data, including both subjective symptom assessments and objective measurements. Understanding these data types theoretically requires knowledge of their variability and how they might impact the trial outcomes.

- **Statistical Nuances:** Theoretical knowledge of statistical methods used in PD trials, such as handling missing data, analyzing longitudinal data, and adjusting for confounding variables, is critical. However, without hands-on experience, philosophizing about these methods can turn out to be ridiculous.

- **Disease Progression and Patient Heterogeneity:** Grasping the nuances of PD's progression and its impact on data collection and interpretation is vital. For instance, the progression rate can vary significantly among patients, affecting data analysis and interpretation.

- **Data Integration Challenges:** PD trials often involve integrating data from various sources, including clinical assessments, patient-reported outcomes, and newer technologies like wearable devices. Practical experience of how to harmonize and analyze these diverse data sets is essential. Theoretical knowledge can only help to ask good questions.

ROLE OF MENTORING IN OVERCOMING THEORETICAL LIMITATIONS

Generative AI, while adept at processing and generating information based on available data, lacks real-world experience and contextual understanding. In the complex field of clinical trials, this lack of experiential learning can lead to gaps in the AI's ability to fully comprehend and address nuanced data challenges regardless of the therapeutic area. This is a perfect example, where human experience clearly outrivals artificial intelligence.

AI, by its current design, follows algorithmic patterns and lacks the creative problem-solving abilities that human researchers possess. In clinical research, where innovative approaches to data challenges are often needed, this can be a significant limitation.

Mentoring by experienced clinical data scientists can bridge the gap between theoretical knowledge and practical application:

- **Real-world Insights**: Experienced professionals can share insights from actual PD trials, highlighting common pitfalls and effective strategies that textbooks may not cover.

- **Case Studies:** Mentors can provide case studies or examples from their experiences, offering a practical perspective that helps in understanding how theoretical concepts are applied in real-world scenarios.

- **Guidance on Methodological Approaches:** Experienced programmers can guide on the selection and implementation of appropriate statistical methods in PD data analysis, considering the specific challenges of the disease.

- **Navigating Complexities**: They can also help in navigating the complexities of PD data, such as understanding the importance of certain variables, interpreting fluctuating symptoms, and making informed decisions about data handling.

In summary, while a theoretical understanding forms the foundation for comprehending data challenges in PD clinical trials, mentorship plays a crucial role in translating this knowledge into practical skills. This combination of theoretical learning and practical insights is vital for anyone looking to contribute effectively to PD research and data analysis.

Nevertheless, the following prompt examples demonstrate the challenge of creating explicit knowledge without having this implicit TA knowledge:

Example of a Poor Prompt
"Write about PD trial data challenges."

Why This is a Poor Prompt:

- **Lacks Specificity:** The prompt doesn't specify that it's about PD or clinical trials, which can lead to generic and irrelevant content.
- **Missing Context:** There's no mention of the need for understanding specific challenges like medication history inaccuracies, disease progression variability, or subjectivity in symptom assessment. It completely disregards the knowledge, which has been created for example in the “Demographics and Baseline Characteristics” section.
- **No Request for Solutions or Methodologies:** It doesn't ask for ways to address these challenges or for the methodologies used in data validation.

Example of a Good Prompt

"Provide a detailed analysis of common data challenges encountered in Parkinson's Disease (PD) clinical trials as described in the attached document. Focus on specific issues such as inaccurate or incomplete medication history, variability in disease progression, subjectivity in symptom assessment, the integration of data from wearable devices, handling missing data, dose-response relationships, and comorbidities.*

Discuss sophisticated statistical methodologies and best practices for addressing these challenges, ensuring data validity, and maintaining the integrity of the study. Highlight the importance of collaboration with clinicians and healthcare professionals in overcoming these obstacles."

**Note*: As you might have realized, this prompt refers to an attached document. In this case, I have utilized an external document, which included a description of the demographic and baseline characteristics. For each of the variables, I already had some initial ideas what might happen during the data collection. The creation of such a document could be an outcome of a learning journey.

Why This is a Better Prompt:

- Detailed and Specific: The prompt clearly specifies the context of PD clinical trials and outlines particular data challenges to be addressed.
- Asks for Methodologies and Solutions: It requests information on how to address these challenges, which is crucial for a comprehensive and practical article.
- Encourages Holistic Understanding: By mentioning collaboration with healthcare professionals, it acknowledges the multidisciplinary approach needed to tackle these issues effectively.

This well-crafted prompt is designed to yield an informative and contextually relevant article, focusing on the intricacies of data challenges in PD clinical trials and providing insights into effective methodologies and collaborative strategies for overcoming these obstacles.

SUMMARY AND CONCLUSION

In the ever-evolving field of clinical data science, possessing a deep understanding of specific therapeutic areas is not just beneficial, it's crucial. However, acquiring this specialized knowledge presents a significant challenge, especially when venturing into unfamiliar domains. This paper has explored the innovative role of generative AI in bridging this knowledge gap, providing clinical data scientists with essential insights into new therapeutic areas, like Parkinson's Disease (PD), which they may not have previously encountered.

Generative AI: A Catalyst for Knowledge Acquisition

Our discussion has illuminated how generative AI can be a powerful tool in the hands of a clinical data scientist seeking to gain expertise in a new therapeutic area. We delved into the effective use of AI-driven prompts, guided by the structured approach recommended by PHUSE educations, to craft informative and educational articles. This methodological approach not only streamlines the learning process but also ensures a comprehensive understanding of the subject matter.

From Foundations to Complexities: The Learning Journey

Following the PHUSE education structure, we've outlined a journey that begins with the basic understanding of a disease and gradually escalates to more intricate aspects like study design, data challenges, and regulatory guidelines. This progression highlights an important observation: while foundational knowledge of a disease might be more straightforward to acquire, the complexity significantly increases when addressing advanced topics such as the nuanced challenges of data management in clinical trials.

AI's Strengths and Limitations: A Balanced View

Throughout this exploration, we've acknowledged both the strengths and limitations of AI in this context. While AI demonstrates clear proficiency in gathering and synthesizing information, it also has its constraints, particularly in areas requiring deep expertise and nuanced understanding. This is where the synergy of AI and human expertise becomes pivotal. The mentorship from seasoned professionals, coupled with AI capabilities, creates a robust framework for knowledge acquisition.

PHUSE Educations: Fostering a Mentoring Ecosystem

PHUSE Educations is committed to supporting this mentorship-driven approach. We encourage aspiring clinical data scientists to take the first step: draft an educational article on a new therapeutic area and submit it for review. Our team of industry veterans stands ready to provide constructive feedback, ensuring that your initial efforts are nurtured into valuable contributions. This collaborative process not only aids individual learning but has the potential to generate a wealth of shared knowledge for the broader community.

In conclusion, the journey of acquiring therapeutic area knowledge, facilitated by AI and enriched through expert mentorship, is a path filled with learning and growth. We at PHUSE Educations are excited to be part of this journey, fostering an environment where knowledge and expertise are shared, benefiting individuals and the community alike.

REFERENCES AND RECOMMENDED READINGS

- PHUSE Educations – Therapeutics Area Cluster: Oncology, Rabia Ahmed (GSK), Alberto Montironi (UCB) (last checked January 9th 2024), https://sway.office.com/JE8JMKzBLI5HF2EZ .
- PHUSE Educations – Therapeutics Area Cluster (last checked January 9th 2024), https://sway.office.com/AcMq5QyFFnWr9khD
- PHUSE Educations - Therapeutics Area – Parkinson's Disease (last checked January 12th): https://sway.cloud.microsoft/aWp3G0HMAsfgBEUI?ref=Link
- CDISC Therapeutic Area Standards, last checked January 9th 2024, https://www.cdisc.org/standards/therapeutic-areas
- CDISC Therapeutic Area Standard for Parkinson's Disease, last checked January 9th 2024,: https://www.cdisc.org/standards/therapeutic-areas/parkinsons-disease

CONTACT INFORMATION

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)
Your comments and questions are valued and encouraged.
Contact the author at:

Author Name Sascha Ahrweiler:
Company: PHUSE LLC
Email: sascha.ahrweiler@phuse.global
Website: www.phuse.global