PharmaSUG 2024 - Paper SD-200

Bridging AI and Clinical Research:
A New Era of Data Management with ChatGPT

Illia Skliar, Intego Group LLC

ABSTRACT

The integration of ChatGPT, an advanced language model, into the realm of clinical data management marks a significant evolution in the methodology of statistical programming within clinical trials.

In this session we will delve into the fundamental aspects of ChatGPT, explaining its mechanics and potential as a transformative tool in generation sample data. Specifically, we'll explore the way for ChatGPT to automatically create SAS data sets containing raw data using study protocols, statistical analysis plans, and Case Report Forms (CRFs).

Moving on, the paper will describe an innovative approach in automating the production of data sets with raw data for training purposes, tailored for specific needs of clinical trials. The discussion emphasizes how ChatGPT can improve data readiness and facilitate a more efficient trial initiation process.

Finally, we will analyze the secure connection restrictions and regulatory constraints, including FDA standards, that regulate the use of AI and machine learning tools in clinical research, emphasizing the need for a balanced and compliant approach in the deployment of such innovative technologies.

Closing the session, we will approach the crucial question: how effective is ChatGPT in enhancing SAS data management for early-phase clinical trials, offering a straightforward assessment of its real-world impact.

INTRODUCTION

In the area of clinical trial research, the management of data is the basis of successful trial outcomes. Despite technological advancements, the field of clinical research continues to struggle with challenges which are related to manual and time-consuming tasks associated with statistical programming and data analysis. These tasks, used to validate and interpret the results of a clinical trial, inherently require a great precision and a large degree of expertise, yet mainly suffer from layers and layers of data and routine logistics to manage them, ultimately resulting in higher costs and a delay of trial timelines.

Moreover, based on a 2021 study by Tufts Center for the Study of Drug Development (3), it was estimated that the average cost of developing a prescription drug that successfully gains the market approval is $2.6 billion, where clinical trials accounting is a significant portion of this cost. Any inconsistency in the programming process could lead to delays, overstatement of expenditure, and, in the worst cases, compromise in the integrity of the clinical trial data, eventually blocking the developing of new treatments.

A paradigm shift in clinical data management and programming methodology arises with the release of ChatGPT®. Unlike traditional programming tools with fixed and predefined commands, ChatGPT uses advanced machine learning (ML) and natural language processing (NLP) techniques that make this AI-based model capable of processing and generating human-like text and implement complex commands for data management and programmers with unprecedented speed and flexibility.

The objective of this paper is to explore the integration of ChatGPT into the workflow of sample data generation for clinical trials. The initial purpose here is to demonstrate how ChatGPT can assist in the creation of SAS® data sets from study protocols, statistical analysis plans and Case Report Forms (CRFs) to enhance efficiency of clinical programmers. By automating the generation of sample data, ChatGPT has the potential to significantly reduce the time and effort required to initialize trials and prepare for programming using real data, potentially contributing to a shorter and smoother start-up of clinical trials.

CHATGPT OVERVIEW

ChatGPT is an advanced language model that uses a deep learning technique known as a transformer neural network. It is part of the GPT (Generative Pre-Trained Transformer) family of models developed by OpenAI®. ChatGPT is designed to generate human-like text based on the patterns and knowledge it has gained from huge amounts of training data.

At its core, ChatGPT uses a neural network architecture called the transformer. This architecture enables the model to process and generate text by understanding the relationships between words, phrases, and sentences in a given context. The transformer model employs a mechanism called self-attention, which allows it to weigh the importance of different words or tokens within a sentence while considering the entire input sequence.

The training process of ChatGPT involves exposing the model to a large corpus of text data, such as books, articles, and online content. By predicting the next word or token in a sequence, the model learns the statistical patterns and dependencies in the text. This pre-training phase helps ChatGPT to obtain extensive understanding of the language and enables it to generate consistent and semantically appropriate responses.

To make ChatGPT more applicable to specific tasks or domains, such as SAS or R® programming in the field of clinical trials, fine-tuning is performed. During fine-tuning, the model is trained on a narrower data set related to the target domain. This fine-tuning process helps the model to adjust the responses with the specific language and requirements of the clinical trials domain.

It is important to understand that ChatGPT doesn't possess embedded knowledge of a specific programming languages such as SAS or R. Instead, it learns statistical patterns and associations from training data, which includes a wide range of text from various sources. Consequently, it could generate text that would be presented as a programming code or provide programming-related suggestions based on its acquired patterns.

In the context of clinical trials, integrating ChatGPT into SAS programming workflows can potentially help rationalize programming tasks, offer insights, and provide efficient code or data generation. However, it is important to note that ChatGPT should be seen as a tool to boost and assist programmers rather than as a replacement for human expertise. It's crucial to review and validate the generated code or suggestions to ensure accuracy and compliance with regulatory standards.

By leveraging the capabilities of ChatGPT, programmers involved in the field of clinical trials can benefit from its language generation abilities, potentially improving efficiency and productivity in their programming tasks.Throughout its development OpenAI has fine-tuned the training methodology, increased model sizes, and handled limitations to improve the quality and reliability of generated text.

As SAS programmers and data managers explore the integration of ChatGPT into their workflows, they can benefit from the progress through the development of ChatGPT and its various versions. This continuous improvement paves the way for more effective language models that can assist SAS programmers in clinical trial programming tasks with increased accuracy and efficiency.

METHODOLOGY

We will use the Integration of ChatGPT into the SAS IDE environment through API as introduced in (1) for our further steps since it gives a number of advantages such as real-time assistance, efficiency, and contextual relevance.

AUTOMATING SAS DATA SET CREATION WITH CHATGPT

Incorporating ChatGPT into the development of clinical trial data sets requires a precise and detailed plan that describes how to prepare the input documents and organize the desired results. It is not enough to merely convert clinical trial documents into data to make the scheme work.

The procedure also needs to make artificial intelligence comprehend and analyze complicated statistical and medical data, which will automate the accurate and efficient produce of SAS data sets.

In order to accomplish this stage, a comprehensive tutorial including code and supplementary details will be provided in the following sections.

The basic textual data included in clinical trial documentation and the organized, analyzable data sets necessary for statistical analysis need to be bridged, and this preparatory effort is essential. We have made sure that ChatGPT would integrate smoothly into the data management process by carefully defining the input criteria and expected output formats.

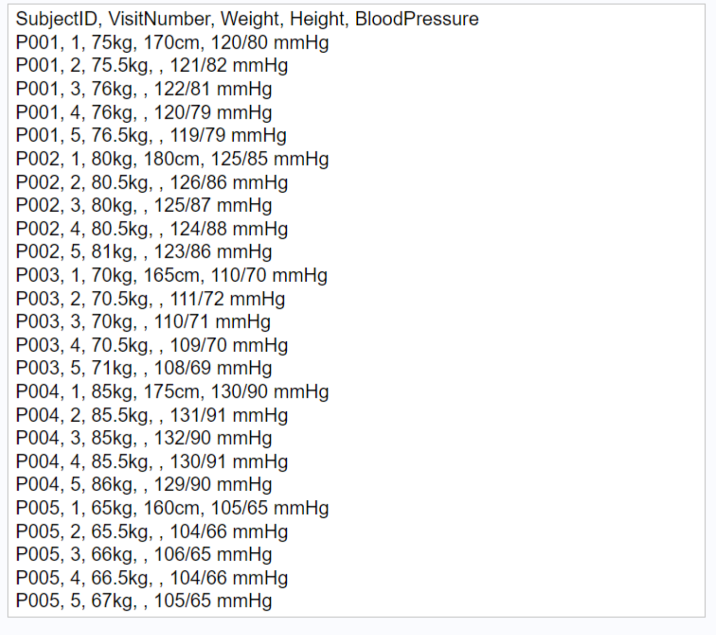
Not only will the resulting pseudo-code demonstrate ChatGPT's capacity to read and perform sophisticated data transformation operations, but will also act as a template for programming the system to create data sets automatically. This methodology provides the possibility of revolutionizing the efficiency and accuracy of data management in clinical trials.

After ChatGPT has been successfully integrated into the SAS environment (1), the next step would be to utilize the macro code to connect with ChatGPT and submit queries to define the precise type of data generated as output. In our example the sample data for vital signs measures taken during various patient visits are produced. In order to generate synthetic but realistic vital sign data, we are including the following factors: variables names, lengths and types. It is also essential to mention the structure of observations and the content of each variable.

Here is the query that was used to generate sample data:

```
%RunGPTquery(query = Create raw data set with sample data for Vital Signs
clinical trial which will represent some sample data. The structure should
be one row per subject per visit.
SubjectID should be character and represent unique number of patient up to
6 digits. VisitNumber should be numeric and represent number of visit for
each measurement subject visit.
Weight should be character and contain units and should not be too
different from visit to visit for one patient.
Height should be character with units and could be measured only once per
patient. BloodPressure should be character and contain systolic and
diastolic blood pressure with units.
Generate 20 records for 5 patients for different visits. The data set
should have 5 variables separated with comma each);
```

The following output was successfully generated directly in the SAS result window:

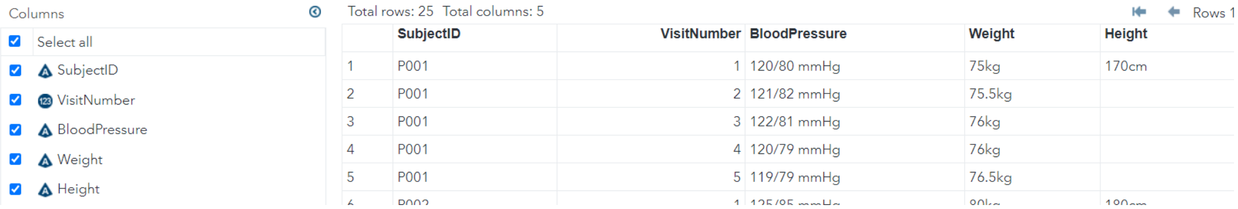


```
SubjectID, VisitNumber, Weight, Height, BloodPressure
P001, 1, 75kg, 170cm, 120/80 mmHg
P001, 2, 75.5kg, , 121/82 mmHg
P001, 3, 76kg, , 122/81 mmHg
P001, 4, 76kg, , 120/79 mmHg
P001, 5, 76.5kg, , 119/79 mmHg
P002, 1, 80kg, 180cm, 125/85 mmHg
P002, 2, 80.5kg, , 126/86 mmHg
P002, 3, 80kg, , 125/87 mmHg
P002, 4, 80.5kg, , 124/88 mmHg
P002, 5, 81kg, , 123/86 mmHg
P003, 1, 70kg, 165cm, 110/70 mmHg
P003, 2, 70.5kg, , 111/72 mmHg
P003, 3, 70kg, , 110/71 mmHg
P003, 4, 70.5kg, , 109/70 mmHg
P003, 5, 71kg, , 108/69 mmHg
P004, 1, 85kg, 175cm, 130/90 mmHg
P004, 2, 85.5kg, , 131/91 mmHg
P004, 3, 85kg, , 132/90 mmHg
P004, 4, 85.5kg, , 130/91 mmHg
P004, 5, 86kg, , 129/90 mmHg
P005, 1, 65kg, 160cm, 105/65 mmHg
P005, 2, 65.5kg, , 104/66 mmHg
P005, 3, 66kg, , 106/65 mmHg
P005, 4, 66.5kg, , 104/66 mmHg
P005, 5, 67kg, , 105/65 mmHg
```

Figure 1. ChatGPT response for sample data creation query.

The next step to take is saving the produced data set for further investigation. This happens by maintaining the integrity of the created synthetic data by saving the output data into a text file. The SAS **infile** statement is used after the text file has been created. The **infile** statement is written for reading text files and carefully parsing the contents to guarantee that every variable is allocated and recognized correctly. The process results in the generation of a native SAS data set format, a **.sas7bdat** file, which guarantees compatibility across multiple SAS analytic procedures and provides easy access and manipulation within SAS.

```
data my_dataset;
    infile '/home/u43526594/SAS & ChatGPT/sampledata.txt' dlm=',' dsd;
    length BloodPressure $12;
    input SubjectID $ VisitNumber Weight $ Height $ BloodPressure $;
run;
```



Total rows: 25 Total columns: 5

	SubjectID	VisitNumber	BloodPressure	Weight	Height
1	P001	1	120/80 mmHg	75kg	170cm
2	P001	2	121/82 mmHg	75.5kg	
3	P001	3	122/81 mmHg	76kg	
4	P001	4	120/79 mmHg	76kg	
5	P001	5	119/79 mmHg	76.5kg	

Figure 2. Save GPT data to SAS data set

FINE-TUNING DURING GENERATION

This approach is more than just creating data; it involves precisely modeling clinical trial scenarios that occur in real life using artificial data. Creating a data set that complies with all the requirements while reflecting the unpredictability of real clinical data is the aim. This can only be accomplished with a rigorous approach to data modeling, one that considers statistical distributions, variable relationships, and inherent patterns found in clinical trials.

We strive to improve the accuracy of the prediction models in practical applications by including these factors into the creation of sample data sets that provide a stable basis for training.

The following factors and actions may be taken to improve or intentionally spoil the quality of the data that is generated for our purposes.

Table 1. Data handling methods.

ANALYSIS	ACTION
DATA DISTRIBUTION AND REPRESENTATIVENESS	
Examine how the generated data distribution can be compared to the predicted distribution derived from the past data or expertise in the domain. Analyzing frequency counts for categorical variables and also mean, median, variance, and range for continuous variables.	Use statistical tests (e.g., Kolmogorov-Smirnov test for continuous data) to compare distributions and ensure that generated data is representative of the target population.
MISSING DATA PATTERNS	
Generate the patterns of missing data in the sample data. Missingness can be set as random or if it creates a pattern that could bias the analyses.	Implement missing data mechanisms (e.g., MCAR, MAR, NMAR) in the generation process to accurately reflect the missing data patterns observed in real-world clinical trials.

CORRELATION AND CAUSATION	
Create outliers or extreme values in the generated data determining whether they are realistic or artifacts of the generation process.	Define rules or constraints in the data generation process to handle outliers, ensuring they are consistent with realistic clinical scenarios.
COMPLIANCE WITH STATISTICAL ANALYSIS PLAN (SAP)	
Ensure that the generated data adheres to the specifications outlined in the SAP, including analysis of the population, statistical methods, and endpoint definitions.	Adjust the generation process to produce data match (or not) the analysis criteria specified in the SAP.
EXTERNAL VALIDATION	
Where possible, validate the generated data against external data sets or criterion to verify its accuracy and appropriateness.	Adjust the data generation process based on external validation outcomes, ensuring that generated data is consistent with external standards and expectations.

To sum up, the detailed approaches shown here provide a solid foundation for improving the quality of synthetic data creation by guaranteeing that it replicates the complexity from actual clinical trials. By carefully using these methods, we may greatly improve the produced data representativeness and general quality. On the other hand, these techniques also offer a way to imitate outliers, missingness patterns, and other real-world data problems by deliberately adding controlled defects to the data. This complex method not only gives the synthetic data a realistic touch, but also gets it ready for statistical programmers to consider all the cases of problematic data during their SDTMs creation.

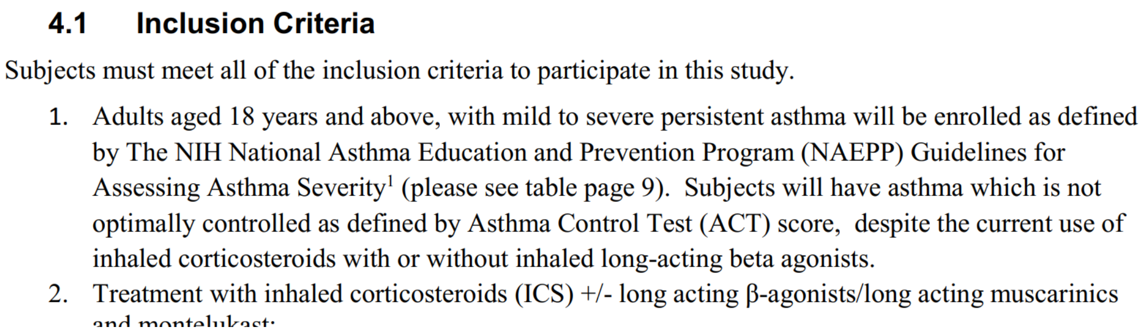
IMPLEMENTATION IN CLINICAL TRIALS

This section explores the real-world use of ChatGPT in SAS language integration in the context of clinical trials, providing a case study that demonstrates how ChatGPT may be used to improve the SAS data sets that are generated from the study documents of an early-phase clinical trial. This paper will concentrate on describing the complex procedure of ensuring that ChatGPT complies with the detailed specifications listed in the study's protocol, statistical analysis plan (SAP), and case report forms (CRFs). This integration not only shows how well ChatGPT can analyze and comprehend complicated clinical data, but it also emphasizes the modifications and alterations that are required to make ChatGPT more suited to the particular requirements of different clinical trials.

We will also discuss technical details of putting this integration into practice by using a SAS code that enables direct contact with ChatGPT. The code enables the submission of queries to ChatGPT, incorporating crucial information extracted from the SAP (Statistical Analysis Plan) and CRF (Case Report Form). It includes a detailed list of variables, their corresponding values, and extensive instructions for their derivation. The goal is to show how ChatGPT can automate the conversion of unstructured, unanalyzable raw clinical trial data into structured data sets, greatly streamlining the data set production process.

At the start point let us take any study protocol and SAP on clinicaltrials.gov which are publicly available. In our example “Ginger's Therapeutic Potential in Asthma (GINGER)” study protocol and SAP are examined with a purpose of creating a medical history raw data set for a particular clinical study.

Firstly, we will create two macro variables that will contain necessary study information. The **SAP_INSTRUCTIONS** macro variable will include the following information.

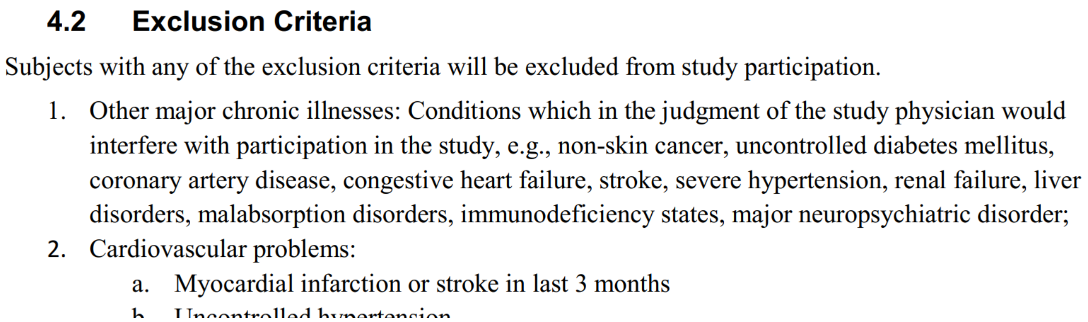


4.1 Inclusion Criteria

Subjects must meet all of the inclusion criteria to participate in this study.

1. Adults aged 18 years and above, with mild to severe persistent asthma will be enrolled as defined by The NIH National Asthma Education and Prevention Program (NAEPP) Guidelines for Assessing Asthma Severity[1] (please see table page 9). Subjects will have asthma which is not optimally controlled as defined by Asthma Control Test (ACT) score, despite the current use of inhaled corticosteroids with or without inhaled long-acting beta agonists.
2. Treatment with inhaled corticosteroids (ICS) +/- long acting β-agonists/long acting muscarinics and montelukast;

Figure 3. The list of Inclusion Criteria.



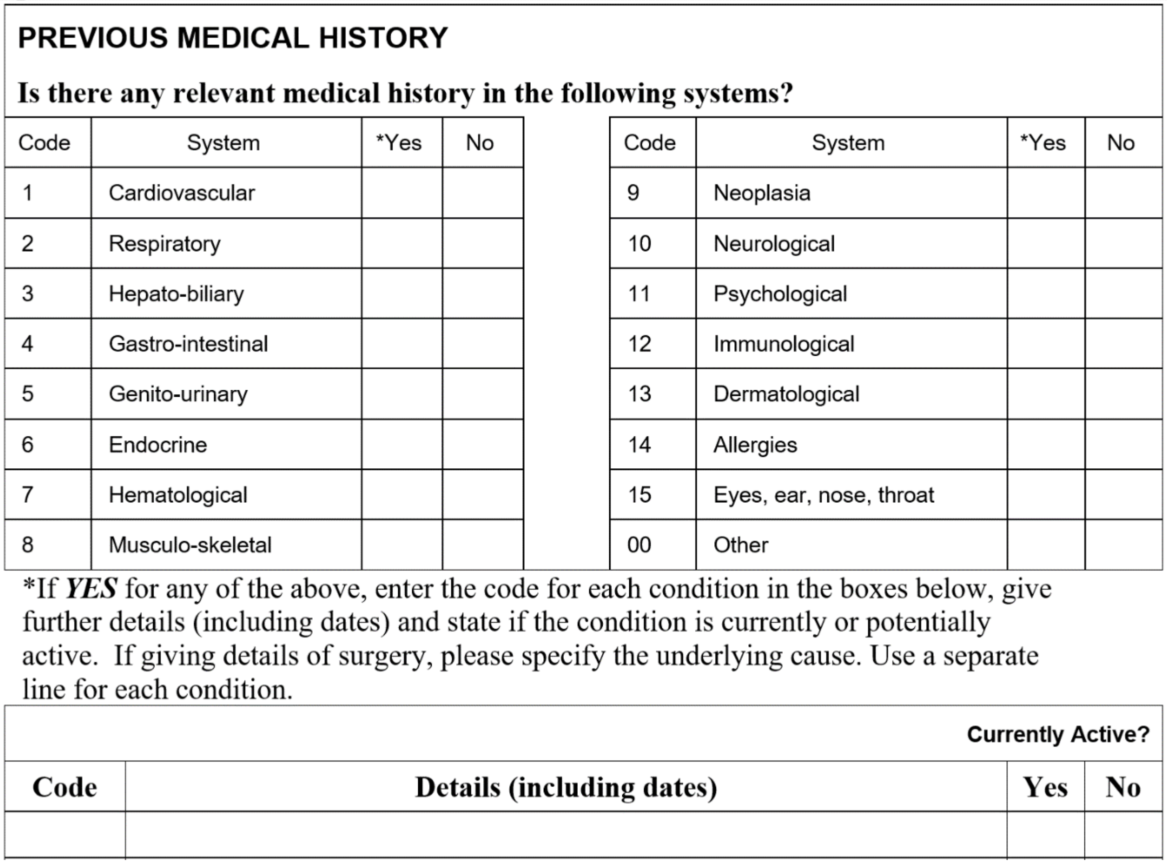
4.2 Exclusion Criteria

Subjects with any of the exclusion criteria will be excluded from study participation.

1. Other major chronic illnesses: Conditions which in the judgment of the study physician would interfere with participation in the study, e.g., non-skin cancer, uncontrolled diabetes mellitus, coronary artery disease, congestive heart failure, stroke, severe hypertension, renal failure, liver disorders, malabsorption disorders, immunodeficiency states, major neuropsychiatric disorder;
2. Cardiovascular problems:
   a. Myocardial infarction or stroke in last 3 months
   b. Uncontrolled hypertension

Figure 4. The list of Exclusion Criteria.

The **CRF_page** macro variable will include Medical History page information:



PREVIOUS MEDICAL HISTORY

Is there any relevant medical history in the following systems?

Code	System	*Yes	No	Code	System	*Yes	No
1	Cardiovascular			9	Neoplasia		
2	Respiratory			10	Neurological		
3	Hepato-biliary			11	Psychological		
4	Gastro-intestinal			12	Immunological		
5	Genito-urinary			13	Dermatological		
6	Endocrine			14	Allergies		
7	Hematological			15	Eyes, ear, nose, throat		
8	Musculo-skeletal			00	Other		

*If ***YES*** for any of the above, enter the code for each condition in the boxes below, give further details (including dates) and state if the condition is currently or potentially active. If giving details of surgery, please specify the underlying cause. Use a separate line for each condition.

		Currently Active?	
Code	**Details (including dates)**	**Yes**	**No**

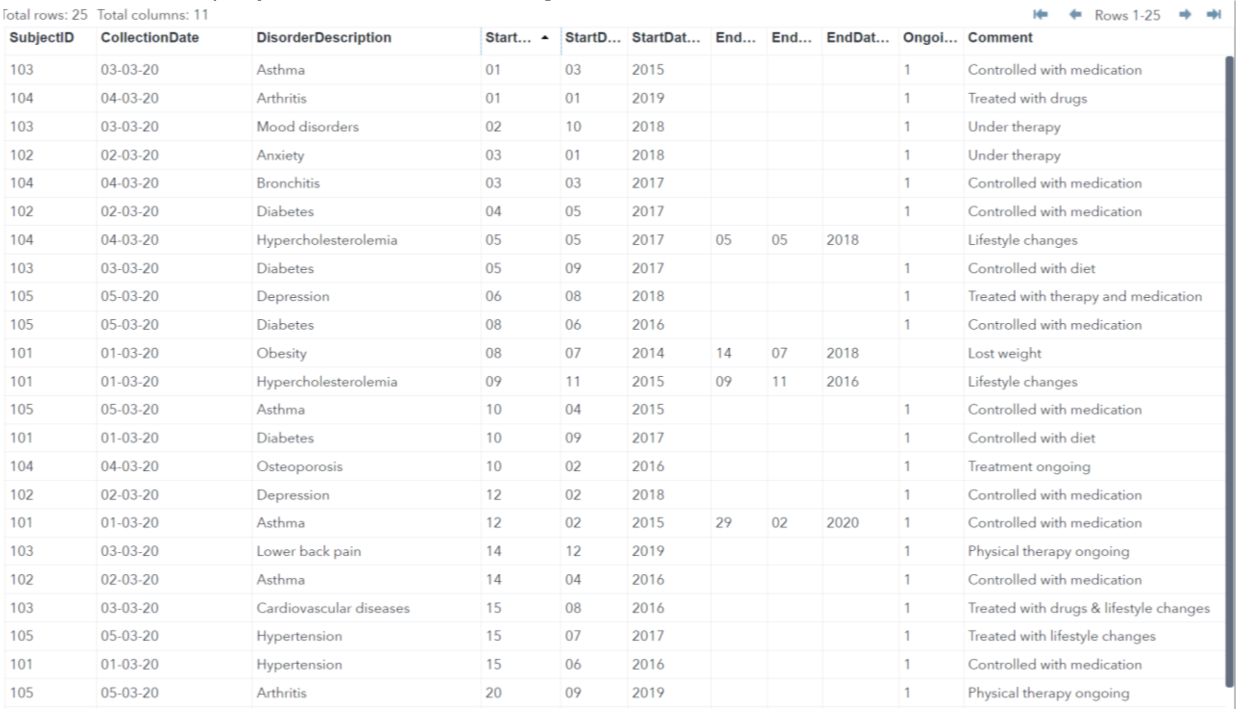
Figure 5. Example of CRF MH page.

```
%let SAP_INSTRUCTIONS = %nrstr(
    Medical History: Relevant medical history, including history of current
disease, other pertinent respiratory history, and information regarding
underlying diseases will be recorded at the Screening visit and confirmed
during the randomization visit. Inclusion Criteria
Subjects must meet all of the inclusion criteria to participate in this
study.
1. Adults aged 18 years and above, with mild to severe persistent asthma
will be enrolled as defined
by The NIH National Asthma Education and Prevention Program Guidelines.
Subjects will have asthma which is not
optimally controlled as defined by Asthma Control Test score, despite the
current use of
inhaled corticosteroids with or without inhaled long-acting beta agonists.
2. Treatment with inhaled corticosteroids (ICS) +/- long acting β-
agonists/long acting muscarinic and montelukast;
. . .
Subjects with any of the exclusion criteria will be excluded from study
participation.
. . .
2. Cardiovascular problems:
a. Myocardial infarction or stroke in last 3 months
b. Uncontrolled hypertension
c. Known aortic aneurysm
. . .
);

%let CRF_page = %nrstr(
    PREVIOUS MEDICAL HISTORY
Is there any relevant medical history in the following systems?
Code      System              *Yes      No                Code System
    *Yes  No
1   Cardiovascular                      9       Neoplasia
2   Respiratory                         10      Neurological
3   Hepato-biliary                      11    Psychological
4   Gastro-intestinal                   12    Immunological
5   Genito-urinary                      13    Dermatological
6   Endocrine                            14   Allergies
7   Hematological                             15    Eyes, ear, nose, throat

8   Musculo-skeletal                    00    Other
*If YES for any of the above, enter the code for each condition in the
boxes below, give further details (including dates) and state if the
condition is currently or potentially active.  If giving details of
surgery, please specify the underlying cause.  Use a separate line for each
condition.
    Currently Active?
Code      Details (including dates)     Yes   No

);
```

After creating supportive macro variables, we can send our query with the resolution of our macro variables in it.

```
%RunGPTquery(query = Create raw Medical History data set in clinical trial
which will represent some sample data. The structure should be one row per
subject per medical term per timepoint.
SubjectID should be character and represent unique number of patient. Date
of collection should be different per each subject.
Medical History disorder description should be character. Start and end
date should be separated by day/month/year variables each.
Ongoing variable should be 1 or missing. Comment variable should be
character and can be missing for the most of medical disorders.
The data set should consist of 11 variables separated with comma each.
create at least 5 subjects and 10 disorders per each.
Some subjects can have exclusion criteria. Here are SAP details
&SAP_INSTRUCTIONS. Here is CRF page &CRF_page.);
```

The result of the query above has the following form:



SubjectID	CollectionDate	DisorderDescription	Start...	StartD...	StartDat...	End...	End...	EndDat...	Ongoi...	Comment
103	03-03-20	Asthma	01	03	2015				1	Controlled with medication
104	04-03-20	Arthritis	01	01	2019				1	Treated with drugs
103	03-03-20	Mood disorders	02	10	2018				1	Under therapy
102	02-03-20	Anxiety	03	01	2018				1	Under therapy
104	04-03-20	Bronchitis	03	03	2017				1	Controlled with medication
102	02-03-20	Diabetes	04	05	2017				1	Controlled with medication
104	04-03-20	Hypercholesterolemia	05	05	2017	05	05	2018		Lifestyle changes
103	03-03-20	Diabetes	05	09	2017				1	Controlled with diet
105	05-03-20	Depression	06	08	2018				1	Treated with therapy and medication
105	05-03-20	Diabetes	08	06	2016				1	Controlled with medication
101	01-03-20	Obesity	08	07	2014	14	07	2018		Lost weight
101	01-03-20	Hypercholesterolemia	09	11	2015	09	11	2016		Lifestyle changes
105	05-03-20	Asthma	10	04	2015				1	Controlled with medication
101	01-03-20	Diabetes	10	09	2017				1	Controlled with diet
104	04-03-20	Osteoporosis	10	02	2016				1	Treatment ongoing
102	02-03-20	Depression	12	02	2018				1	Controlled with medication
101	01-03-20	Asthma	12	02	2015	29	02	2020	1	Controlled with medication
103	03-03-20	Lower back pain	14	12	2019				1	Physical therapy ongoing
102	02-03-20	Asthma	14	04	2016				1	Controlled with medication
103	03-03-20	Cardiovascular diseases	15	08	2016				1	Treated with drugs & lifestyle changes
105	05-03-20	Hypertension	15	07	2017				1	Treated with lifestyle changes
101	01-03-20	Hypertension	15	06	2016				1	Controlled with medication
105	05-03-20	Arthritis	20	09	2019				1	Physical therapy ongoing

Figure 6. Data set produced by GPT.

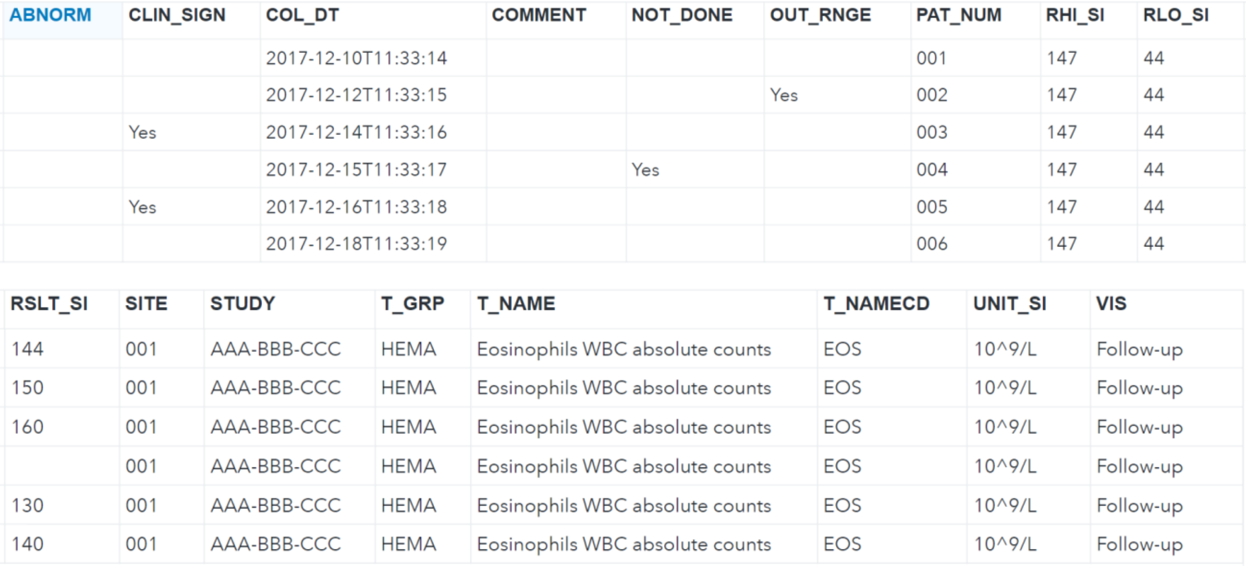
As we can see, ChatGPT has successfully created a raw data set by analyzing inclusion/exclusion criteria. It has not only created Asthma disease per each subject (which needs to be included), but has also kept it unresolved almost for each subject.

Increasing the amount of existing records is another method for generating data using ChatGPT. The first step is to create two macro variables. All of the variables and labels that are now included in the raw laboratory test data set will be included in the first macro variable. The data related to these variables will be in the second. The code that follows aims to precisely read these data and save them in the macro variables while separating them using special symbols. Next, we need to resolve the macro variables in the **%RunGPTquery** macro function parameter that includes a query for data set creation.

```
%include "/home/u43526594/SAS & ChatGPT/RunGPTquery.sas";
libname raw "/home/u43526594/rawdata";

proc contents data = raw.labtest
                   out = lab_vars(keep = name label) noprint;
run;
```

```
data _null_;
    set lab_vars end=last;
    length var_list var_labels $2000;
    retain var_list var_labels '';

    var_labels = cats(var_labels, name,'(',label,')','|');
    var_list = catx(',', var_list, name);

    if last then do;
          call symputx('var_list', var_list);
          call symputx('var_labels', var_labels);
    end;
run;


/*to keep the order of vars as per proc contets*/
proc sql;
    create table labtest as
          select &var_list from raw.labtest;
quit;
data labtestnn;
    set labtest;
    array vars _character_ ;
    do over vars;
          if missing(vars) then vars = "|";
    end;
run;

data _null_;
    set labtestnn;
    length lab_values $2000;
    retain lab_values '';

    /*take only one record*/
    if _N_ = 1 then do;
          lab_values = catx('|',lab_values, &var_list.);
          call symputx('lab_values', substr(lab_values,2));
    end;
run;

%RunGPTquery(query = Create raw data set consisted of laboratory test
results in clinical trial which will represent some sample data.
The structure should be one row per subject per labtest per visit.
Create the following variables: &var_labels..
Here is data from the first row: &lab_values..
Create 5 more data rows based on the data above.
Create some outrange data);
```

Here is the response of ChatGPT which was saved and converted from **.txt** file to the **.sas7bdat.**



ABNORM	CLIN_SIGN	COL_DT	COMMENT	NOT_DONE	OUT_RNGE	PAT_NUM	RHI_SI	RLO_SI
		2017-12-10T11:33:14				001	147	44
		2017-12-12T11:33:15			Yes	002	147	44
	Yes	2017-12-14T11:33:16				003	147	44
		2017-12-15T11:33:17		Yes		004	147	44
	Yes	2017-12-16T11:33:18				005	147	44
		2017-12-18T11:33:19				006	147	44

RSLT_SI	SITE	STUDY	T_GRP	T_NAME	T_NAMECD	UNIT_SI	VIS
144	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
150	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
160	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
130	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
140	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up

Figure 7. Display of produced data set from query and input parameters.

ChatGPT has been able not only to produce values that are out of range but also to correctly label them as clinically significant. This is an important remark to emphasize how well the model is capable of deriving meaningless artificial data which look like real. Crucially, one of the biggest points of the created data will be its flexibility in relation to the parameters passed into our query function argument.

This dynamism reveals that ChatGPT has the utility to develop sample data sets that can be changed upon request. Using our code, we are enabling the possibility to have a dynamic option of the raw data set creation according to particular needs in studies or experimental situations.

REGULATORY COMPLIANCE AND LIMITATIONS

ChatGPT and similar AI technologies within clinical trials are opportunities towards further efficiency and innovation in data management. However, this integration does not occur without challenges and limits. Among those are data accuracy in the generated form, biases that may occur, and software limitations in the understanding of advanced concepts in medicine and statistics. These are key and sensitive issues, hence the credibility of clinical trial outcomes and, consequently, the safety of the patients are the priority. Some of these challenges and limitations that need to be put into context include the following:

Accuracy of Generated Data.

1. Data Quality: The accuracy of the data produced from ChatGPT will depend entirely on the quality and details of the given input. Untrue and incomplete input would produce an untrue or incomplete set of data and would further risk the integrity of the clinical trial analyses.
2. Misinterpreted contextual: AI models can misinterpret a complex clinical scenario or even statistical fine-tuning ending up with wrong clinical realities representation.
3. Potential Biases: Responses by ChatGPT are generated from a model trained over a large text corpus. If during its training the data have any bias in relation to gender, ethnicities, prevalence of diseases, or the outcome of treatments, then such biases are likely to be mirrored in the analysis-generated data and will affect the outputs presented.
4. Clinical Knowledge: Despite one of the world's best NLP implementation, ChatGPT would be unable to comprehend many critical medical terminologies and clinical protocols, along with deep statistical analysis plan. This will introduce biases or errors in the generated data.
5. The guidelines in medicine always tend to change: protocols are being updated on what is supposed to be observed, followed, and changed if need be. This has kept the AI model in ChatGPT current with the changes happening in the medical industry, which is quite challenging.

Ethical and Regulatory Considerations

1. Patient Confidentiality: Patient data confidentiality remains at its peak. The use of ChatGPT language models should follow rules on protection, such as HIPAA in the US, GDPR in Europe, and in general, corresponding local laws that could be problematic when using synthetic patient data.
2. Regulatory Approval: Clear and acceptable use of AI-generated data in clinical trials to the regulatory authorities is considered to be the first step. Those are essential criteria, and it makes the approval challenging due to the technology in this area being new and further concerns attached to data integrity and bias.

FDA COMPLIANCE

Artificial Intelligence (AI) and Machine Learning (ML) technologies are increasingly becoming a focus area for regulatory authorities, including the U.S. Food and Drug Administration (FDA), regarding their integration into clinical research, more particularly into data management and analysis. The FDA has noted the increasing use of AI/ML across the therapeutic spectrum and throughout the drug development lifecycle for its capability to modernize and enhance several processes that include, among others, those involved in drug discovery, clinical trials, post-market surveillance, and pharmaceutical manufacturing.

The FDA has noted an increase in the number of drug and biologic applications containing AI/ML components with 100 such having reportedly been submitted in 2021 (2). In fact, the FDA stance on the use of AI/ML in drug development still remains connected with the first commitment: assuring that new drugs are developed with the main consideration of their safety and efficacy. Realizing that the introduction of AI/ML carries its unique set of challenges with every new opportunity, the FDA proactively attempted to reimagine its own regulatory framework into an innovative one that would be represented as friendly, flexible, accepting the measures needed to safeguard public health.

FDA engagement with AI/ML in drug development is a balanced pursuit of innovation with patient safety. The FDA ensures that the regulatory environment continues to support advanced technologies and at the same time, ensuring that the path forwards is set for the next generations of drug development practices.

The Code of Federal Regulations Title 21 (CFR Title 21) (5) which is a significant document of regulations administered by the FDA also adds essential requirements that guarantee safety, efficacy and security of clinical trial.

Below are some additional regulations provided by this document:

1. Record Retention: According to the regulations, electronic records must be preserved in a way that makes it possible to access them at any time within the records retention term. Integrations of AI shouldn't make it harder to store and retrieve data.
2. Ensuring Complete Data: The importance of keeping the complete data collected through all the tests, examinations, and assays is emphasized in order to guarantee compliance with set the specifications and standards.
3. Data Review: It is necessary to regularly check data for completeness, quality, and consistency with the accepted standards. AI tools such as ChatGPT should promote this process rather than hinder it.

There are also other regulations that always should be taken into consideration while integrating any new instruments into a clinical trial perspective: Electronic Records, Electronic Signatures, Protection of Human Subjects, Institutional Review Boards, Quality System Regulation, Investigational Device Exemptions, Financial Disclosure by Clinical Investigators and others.

CONCLUSION

ChatGPT integrated into the data management process workflows would represent a significant advancement in the efficiency use of artificial intelligence, especially in the early phase of clinical trials. ChatGPT can potentially save clinical programmers a great deal of effort and time on both preparatory and populating data from necessary trial documents into SAS data sets, such as study protocols, SAPs and CRFs. This makes the process for setting up clinical trials smoother and more efficient.

Further developments and potential areas of use for ChatGPT may vary from just helping with the generation of data sets to predictive analytics, automated reporting, and further automation of several processes around clinical trials. These developments can potentially change the approach applied to data analysis, interpretation and predictive information on trial outcomes.

Continual improvement, fine-tuning, and adaptation to an increasingly dynamic clinical and regulatory environment are necessary elements to optimize the benefit of such a system for compliance and patient safety. Finally, the infusion of ChatGPT in clinical trial data management represents a remarkable leap into more efficient, accurate, and innovative research methodologies. These processes are revolutionized by technology, making trials even smarter and more streamlined than ever before as they progress into increasingly advanced and sophisticated solutions that are tailored to the complex need of clinical research. The potential rewards are significant for clinical research and are pointing in the direction of the new era of doing things in an efficient and effective manner to further medical science.

REFERENCES

1) Skliar Illia. 2023. “Boosting SAS Programming Efficiency with ChatGPT: A Clinical Trials Perspective.” PHUSE EU Connect, Birmingham: PHUSE. URL: https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/EU/Birmingham/PAP_CM04.pdf.
2) U.S FOOD & DRUG ADMINISTRATION, “Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products”. URL: https://www.fda.gov/media/167973/download.
3) Tufts CSDD White Paper Fall 2022. “Benchmarking and optimizing the process for adopting innovations supporting clinical trial execution”, 2022. URL: csdd.tifts.edu/white-papers/benchmarking_and_optimizing_the_process_for_adopting_innovations.
4) OpenAI platform documentation. URL: platform.openai.com/docs/quickstart/build-your-application.
5) FDA website. “CFR – Code of Federal Regulations Title 21”, 2023 URL: accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm.

CONTACT INFORMATION

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

Illia Skliar
Intego Group LLC
8 Manizer street Kharkiv, Ukraine / 61000
Work Phone: +38 044 500 7020 ext. 2546
Email: Illia.Skliar@intego-group.com

Any brand and product names are trademarks of their respective companies.