Paper ML14

AI Got the Power: Streamlining Clinical Data Creation

Emily Yates, Formation Bio, NYC, USA
Matt Luppino, Formation Bio, NYC, USA
Andrew Burd, Formation Bio, NYC, USA

ABSTRACT

Access to realistic test Electronic Data Capture (EDC) data is a critical bottleneck for statistical programming and clinical analytics teams. Traditional methods of test data creation rely on manual data entry of just a handful of subjects, which is time-consuming, labor-intensive, and ultimately creates an insufficient amount of data. This presentation will explore how AI-driven solutions, such as ChatGPT and Mockaroo, can enhance the efficiency and quality of test data generation. By reducing the time required for test data creation, programming can commence earlier, shortening deliverable timelines in clinical studies. Improved test data quality also supports the development of more reliable code, minimizing coding errors and reducing the need for rework during studies. This presentation underscores the importance of adopting AI-driven solutions to address the challenges of generating realistic test data in clinical trials.

INTRODUCTION

In the rapidly evolving landscape of clinical trials, the efficiency and quality of data management play a crucial role in accelerating drug development and bringing life-saving therapies to patients. One of the most significant bottlenecks in this process is the creation of realistic test data for Electronic Data Capture (EDC) systems. This data is essential for statistical programming and clinical analytics teams to develop and validate their analysis tools and processes.

Traditionally, test data creation has relied on manual data entry for a limited number of test subjects. This approach is not only time-consuming and labor-intensive but also produces an insufficient amount of data to thoroughly test all possible scenarios. The limitations of this method have far-reaching consequences, including:

1. Delayed start to programming activities
2. Increased risk of coding errors due to inadequate test coverage
3. Extended timelines for study deliverables
4. Potential compromises in patient safety due to insufficient data validation

At Formation Bio, our AI and tech enabled data management, clinical analytics, and engineering teams have been exploring innovative solutions to address these issues and improve the overall efficiency of our clinical trials. This paper examines the potential of AI-driven solutions, such as large language models (LLMs) like ChatGPT and data generation tools like Mockaroo, to revolutionize the test data generation process in clinical trials. By leveraging these technologies, we can significantly reduce the time required for test data creation, allowing programming to begin earlier and shortening deliverable timelines. Moreover, the improved quality and diversity of AI-generated test data support the development of more robust code, minimizing errors and reducing the need for rework during studies.

In the following sections, we will explore the current industry standard practices for test data generation, their limitations, and the value proposition of adopting AI-driven solutions. We will also discuss the specific use cases for high-quality test data and provide insights into the implementation of these technologies in clinical trial settings.
As the pharmaceutical industry continues to embrace digital transformation, the adoption of AI-driven solutions for test data generation represents a significant step towards more efficient, reliable, and innovative clinical trials. This paper aims to demonstrate how these technologies can address the long-standing challenges of generating realistic test data, ultimately contributing to faster and more cost-effective drug development processes.

LIMITATIONS OF INDUSTRY STANDARD PRACTICE

The generation of test data for clinical trials is a critical process that has long relied on manual methods. These practices, while established, often fall short of meeting the evolving needs of modern clinical research. To understand the significance of AI-driven solutions, it's essential to first examine the current industry standard practices and their limitations.

STANDARD TEST DATA GENERATION PROCESS

The typical process for generating test data in clinical trials follows a structured approach:

1. **Prerequisite Documentation**: Test data generation does not commence until the following documents are finalized:
   - Final study protocol
   - Final approved Case Report Form (CRF)

    - Final annotated CRF
    - Final database specifications
    - CRF completion guidelines
2. **Manual Data Entry**: Data is manually entered into the Electronic Data Capture (EDC) system, typically using only one test site.
3. **Limited Test Cases**: A small set of test cases is created, usually comprising 8-10 subjects. These typically include:
    - Screen fail subject
    - Complete subject
    - End of study (EOS) subject
    - Early termination subject
    - Termination due to death
    - Additional test subjects of interest per protocol design
    - Two subjects per scenario per treatment arm
4. **Timeframe**: According to industry sources, such as ICON[1], the standard timeframe for generating test data is approximately three weeks.

BENEFITS OF HIGH-QUALITY TEST DATA

Limitations of Current Practices	Benefits of High-Quality Test Data
Resource Intensive: The manual nature of data entry is highly labor-intensive and costly, requiring significant time and human resources. Additionally, the limited number of test cases is often inadequate to guarantee high-quality downstream programming. This frequently leads to reworking of statistical programming after dry runs, which is time-consuming and costly.	**Time and Cost Efficient:** High-quality test data significantly reduces the number of iterations required in programming, leading to substantial time savings. It enables an earlier start to programming activities, allowing statistical programming and monitoring efforts to begin during Study Start-Up (SSU) and even before CRFs are finalized.
Limited Scope and Flexibility: The standard test cases typically cover only basic scenarios, leaving many visits, edge cases, and potential data issues unexplored. This can result in unexpected data surprises later in the trial process. Manually entered test data is not easily adaptable to changes in specifications. When modifications are required, the data often needs to be re-entered, leading to delays and increased workload.	**Flexibility and Innovation:** High-quality test data provides crucial support for implementing and testing adaptive trial designs. It allows teams to validate complex, dynamic study structures before live data collection begins. Furthermore, comprehensive test data enables more effective setup and validation of risk-based quality management (RBQM) strategies, fostering innovative approaches to study monitoring and data management.
Inadequate Scenario Coverage: Even for its intended purposes, such as statistical programming and EDC UAT, the quality and coverage of the test data are often suboptimal.	**Risk Mitigation:** High quality test data enables thorough testing of all study processes before live data collection allowing teams to identify and address potential issues in data collection, management, and analysis processes, significantly reducing the risk of major problems arising during the live study.

UNLOCKING NEW CAPABILITIES

High-quality test data not only improves existing processes but also enables new capabilities:

1. **Enhanced Medical and Safety Monitoring**: Realistic test data allows for better setup and validation of medical review and safety monitoring processes.
2. **Protocol Deviation Management**: Improved test data facilitates more thorough testing of protocol deviation detection and management systems.
3. **Vendor Data Reconciliation**: Comprehensive test data allows for early testing and refinement of vendor data reconciliation processes.
4. **Early TFLs**: With comprehensive test data, Tables, Figures, and Listings (TFLs) can be developed and refined earlier in the study process.

THE RIPPLE EFFECT ON STUDY QUALITY

The value of high-quality test data extends beyond immediate operational benefits. It contributes to an overall improvement in study quality by:

[1] https://phuse.s3.eu-central-1.amazonaws.com/Archive/2018/Connect/EU/Frankfurt/PAP_PP11.pdf

- Enabling more thorough validation of study processes before live data collection with real patients begins
- Providing a solid foundation for data-driven decision-making throughout the study lifecycle
- Enhancing the ability to identify and mitigate risks early in the study process

In essence, investing in high-quality test data represents a paradigm shift from a reactive to a proactive approach in clinical trial data management. It allows teams to anticipate and address potential issues before they become problems, ultimately leading to more efficient, reliable, and innovative clinical trials.

DEFINING HIGH QUALITY TEST DATA IN CLINICAL TRIALS

Before delving into AI-driven solutions, it's crucial to establish what constitutes high-quality test data in the context of clinical trials. High-quality test data should not only meet the basic requirements of data completeness and accuracy but also provide comprehensive coverage of potential scenarios and edge cases that may occur during a real clinical trial.

CHARACTERISTICS OF HIGH-QUALITY TEST DATA

1. **Comprehensiveness**: High-quality test data covers all pages and domains in the Case Report Form (CRF). It includes data for all scheduled assessments in complete subject profiles and represents various subject scenarios.
2. **Realism**: The data reflects real-world patient characteristics and disease progression. It includes logical relationships between different data points, such as consistent vital signs and lab values within expected ranges for the condition being studied.
3. **Variability:** High-quality test data incorporates both partial and complete dates to reflect real-world data collection challenges. It includes unscheduled visits and assessments, and represents different treatment arms and randomization scenarios to cover all aspects of the trial design.
4. **Edge Cases and Protocol Non-Compliance:** The data includes test cases for common protocol deviations, such as out-of-window visits and missed assessments. It also represents scenarios like screen failures who receive treatment or subjects randomized despite violating inclusion/exclusion criteria, to test the robustness of conditional forms and fields and analysis processes.
5. **Data Inconsistencies and Errors:** High-quality test data incorporates realistic data entry errors, such as typos and inconsistent units. It includes scenarios with missing or incomplete data to mimic real-world data collection challenges and test data cleaning procedures.
6. **Longitudinal Consistency:** The data maintains a logical progression of dates across visits and reflects expected changes in clinical measures over time. This ensures that the test data can effectively validate time-dependent analyses and reports.
7. **Coding Accuracy:** High-quality test data includes accurately coded adverse events, medical history, and concomitant medications. It represents various coding scenarios, including partially coded and uncoded terms, to test the full range of medical coding processes.
8. **Integration Challenges:** The data simulates scenarios that test integration with external data sources, such as central labs and imaging data. It includes cases that may fail vendor data reconciliation to ensure robust testing of data integration processes.
9. **Volume and Scalability:** High-quality test data provides a sufficient volume of data to stress-test systems and processes. It scales to represent multi-site, multi-country trials when necessary, ensuring that data management systems can handle the full scope of the planned study.
10. **Efficiency and Adaptability:** High-quality test data can be generated quickly and efficiently, significantly reducing the time required for study start-up activities. The data generation process is flexible and can easily accommodate changes to the EDC build or study protocol. This adaptability allows for rapid updates to the test data set when modifications are made to CRFs, edit checks, or other study parameters, ensuring that the test data always aligns with the current study design without requiring extensive manual rework.

By meeting these criteria, high-quality test data provides a robust foundation for validating study design, data management processes, and statistical analysis plans. It enables thorough testing of electronic data capture systems, edit checks, and data cleaning processes, ultimately contributing to higher data quality in the actual clinical trial. The challenge lies in generating such comprehensive and nuanced test data efficiently and at scale.

COMPARATIVE ANALYSIS OF TEST DATA GENERATION SOLUTIONS

In the quest for high-quality test data, various approaches have emerged, each with its own strengths and limitations. This section provides a detailed examination of different solutions for test data generation in clinical trials, analyzing their pros and cons to offer a comprehensive understanding of the landscape. The solutions are presented in increasing technical sophistication.

SOLUTION 1: EDC-BASED MANUAL DATA ENTRY

METHOD

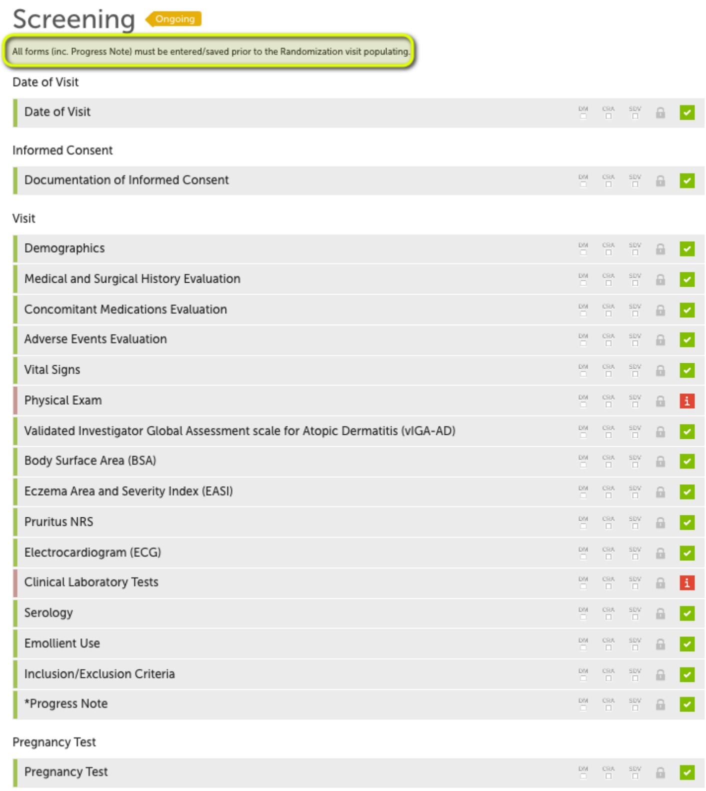
The traditional approach to generating test data involves allocating additional human resources to manually enter data into the Electronic Data Capture (EDC) system. This method typically engages members of the Clinical Data Management (CDM) team or contractors to input data directly into the EDC, mirroring the process used in live trials.

STRENGTHS

1. **Simplest Solution:** Since there are already CDMs and study team members familiar with the protocol, database, and data model, no training is required to implement this solution. Test data can be entered as a part of UAT to simultaneously support UAT of the EDC build and downstream programming teams.
2. **Structural Accuracy and Form Interactions:** EDC-based manual data entry creates data with unparalleled structural accuracy, maintaining consistency in variable names, data types, and overall structure. The resulting dataset precisely reflects the EDC's architecture, including the intricate relationships between different forms and the gating logic built into the system.

WEAKNESSES

1. **Time-Intensive and Resource-Heavy:** The process is extremely time-consuming, often requiring weeks to generate a comprehensive dataset, which can strain project timelines. It's highly resource-intensive, necessitating a team of trained personnel to accomplish the task in a reasonable timeframe. The resource demands can be particularly challenging for smaller organizations or studies with limited personnel.
2. **Inflexibility to Changes:** The method lacks flexibility in the face of changes to the EDC build. When CRFs are changed during UAT and SSU that requires another round of manual data entry for any downstream use cases. The inflexibility can lead to considerable delays and increased workload when changes occur, impacting the overall study timeline and budget.



Screening Ongoing

All forms (inc. Progress Note) must be entered/saved prior to the Randomization visit populating.

Date of Visit

- Date of Visit ✓

Informed Consent

- Documentation of Informed Consent ✓

Visit

- Demographics ✓
- Medical and Surgical History Evaluation ✓
- Concomitant Medications Evaluation ✓
- Adverse Events Evaluation ✓
- Vital Signs ✓
- Physical Exam ℹ
- Validated Investigator Global Assessment scale for Atopic Dermatitis (vIGA-AD) ✓
- Body Surface Area (BSA) ✓
- Eczema Area and Severity Index (EASI) ✓
- Pruritus NRS ✓
- Electrocardiogram (ECG) ✓
- Clinical Laboratory Tests ℹ
- Serology ✓
- Emollient Use ✓
- Inclusion/Exclusion Criteria ✓
- *Progress Note ✓

Pregnancy Test

- Pregnancy Test ✓

Figure 1: An EDC showing the large number of forms needed to be entered to get a subject just to the randomization form

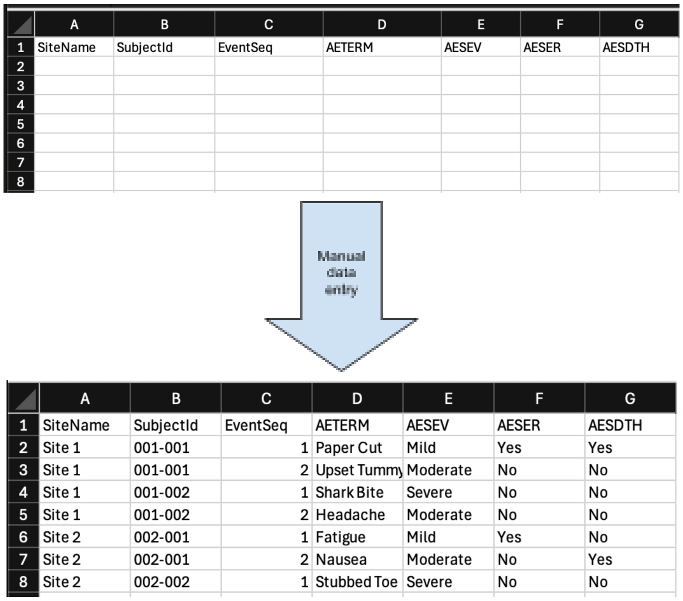
SUMMARY

Manual data entry into the EDC is the simplest method of test data generation and provides the most authentic representation of trial data. However, its significant time and resource cost, coupled with its inflexibility in the face of study changes, make it less suitable for the iterative and focused needs of statistical programming and data analysis.

SOLUTION 2: SPREADSHEET-BASED GENERATION

METHOD

Spreadsheet-based generation involves exporting an empty table structure from the EDC and manually filling in the data using software like Microsoft Excel. This export includes column names, data types, and potentially some metadata about form structure. This method is likely used by a programmer to make a limited dataset for the specific purpose of writing a script. While the export from the EDC will have all of the expected columns, the programmer may choose to only fill in the necessary columns used in the programming to limit the amount of test data generation needed. This method offers a middle ground between the structural accuracy of EDC-based entry and the flexibility of custom data creation.



	A	B	C	D	E	F	G
1	SiteName	SubjectId	EventSeq	AETERM	AESEV	AESER	AESDTH
2							
3							
4							
5							
6							
7							
8							

	A	B	C	D	E	F	G
1	SiteName	SubjectId	EventSeq	AETERM	AESEV	AESER	AESDTH
2	Site 1	001-001	1	Paper Cut	Mild	Yes	Yes
3	Site 1	001-001	2	Upset Tummy	Moderate	No	No
4	Site 1	001-002	1	Shark Bite	Severe	No	No
5	Site 1	001-002	2	Headache	Moderate	No	No
6	Site 2	002-001	1	Fatigue	Mild	Yes	No
7	Site 2	002-001	2	Nausea	Moderate	No	Yes
8	Site 2	002-002	1	Stubbed Toe	Severe	No	No

Figure 2: Export of a selection of columns from the EDC are manually filled in to create test data for programming

STRENGTHS

1. **Time-Efficient:** Spreadsheet-based generation excels in its ability to quickly create focused datasets for specific programming needs, particularly when multiple subjects or specific scenarios are required. Generating data for specific forms, like the end-of-study form, for multiple subjects can be accomplished more efficiently than creating full subject profiles in the EDC.
2. **Customization for Edge Cases:** The method's flexibility enables the manual addition of edge cases to support particular programming requirements, enhancing the robustness of testing scenarios.

WEAKNESSES

1. **Risk of Human Error:** This approach requires intimate knowledge of the EDC build to ensure data accuracy and relevance. Without direct integration with the EDC system, the method doesn't inherently reflect the EDC's gating logic, potentially allowing for impossible data combinations. The manual entry will not enforce the specific codelists embedded in the EDC, leading to potential inconsistencies (e.g., varying formats for sex: m/f, M/F, Male/Female, 1/0).
2. **Inflexibility to EDC Changes:** As the clinical trial evolves and the EDC structure is modified, the spreadsheet-based data may quickly become outdated, requiring manual updates. Aligning the spreadsheet data with EDC changes can be time-consuming and prone to errors, potentially leading to misalignments between the test data and the current EDC configuration.

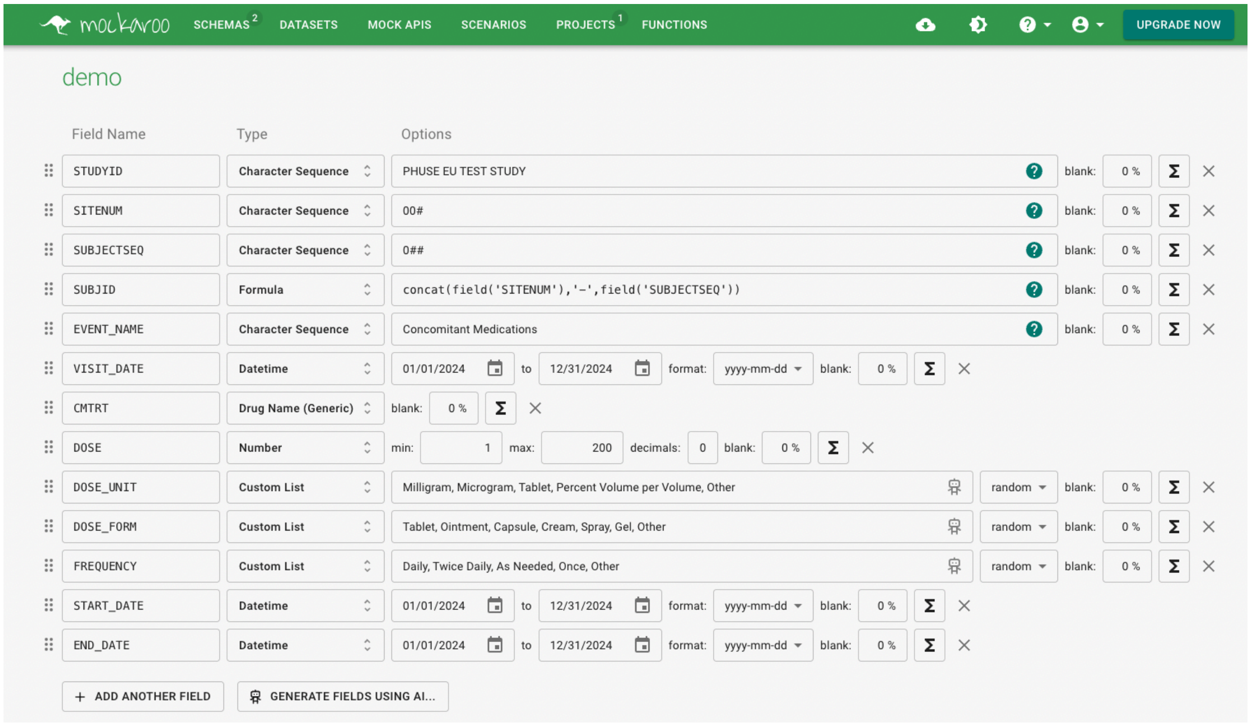
SUMMARY

Spreadsheet-based generation offers a valuable compromise between speed and customization, making it particularly well-suited for the iterative needs of statistical programming and data analytics tasks. Its ability to quickly generate focused datasets and incorporate specific test scenarios provides significant advantages over the more rigid method of entering data directly into the EDC. However, the trade-off comes from potential inconsistencies with the actual EDC system. While this approach can significantly accelerate certain aspects of test data creation, it requires careful management to ensure data integrity and alignment with the evolving EDC structure.

SOLUTION 3: MOCK DATA GENERATION WITH MOCKAROO

METHOD

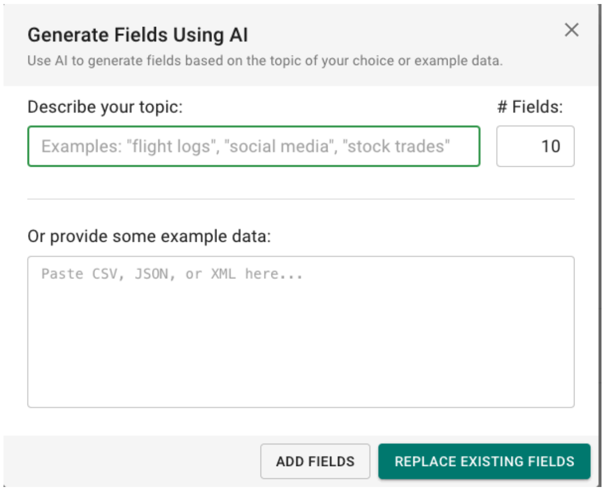
Mockaroo is a tool that generates large volumes of realistic test data based on specified parameters. The method involves providing Mockaroo with column names, data types, codelists, and ranges of values. It can incorporate missing data at user-defined percentages and supports a wide variety of data types, including custom lists, datetime, boolean, drug names, and ICD9/ICD10 codes. Users have the option to build the configuration manually from scratch or use AI to assist in this process. You can use AI by passing some example data to Mockaroo, such as an export from the EDC with a few rows filled in. Mockaroo will then use AI to guess data types, codelists, and ranges if provided with a few rows of test data. This significantly reduces the configuration time. It also employs AI to generate complex column types and derivations based on the user's text input. For reference, putting together this demo took me less than 30 minutes to complete.



demo

Field Name	Type	Options
STUDYID	Character Sequence	PHUSE EU TEST STUDY blank: 0 %
SITENUM	Character Sequence	00# blank: 0 %
SUBJECTSEQ	Character Sequence	0## blank: 0 %
SUBJID	Formula	concat(field('SITENUM'),'-',field('SUBJECTSEQ')) blank: 0 %
EVENT_NAME	Character Sequence	Concomitant Medications blank: 0 %
VISIT_DATE	Datetime	01/01/2024 to 12/31/2024 format: yyyy-mm-dd blank: 0 %
CMTRT	Drug Name (Generic)	blank: 0 %
DOSE	Number	min: 1 max: 200 decimals: 0 blank: 0 %
DOSE_UNIT	Custom List	Milligram, Microgram, Tablet, Percent Volume per Volume, Other random blank: 0 %
DOSE_FORM	Custom List	Tablet, Ointment, Capsule, Cream, Spray, Gel, Other random blank: 0 %
FREQUENCY	Custom List	Daily, Twice Daily, As Needed, Once, Other random blank: 0 %
START_DATE	Datetime	01/01/2024 to 12/31/2024 format: yyyy-mm-dd blank: 0 %
END_DATE	Datetime	01/01/2024 to 12/31/2024 format: yyyy-mm-dd blank: 0 %

Figure 3: Mockaroo's configuration page showing multiple data types and options



Generate Fields Using AI

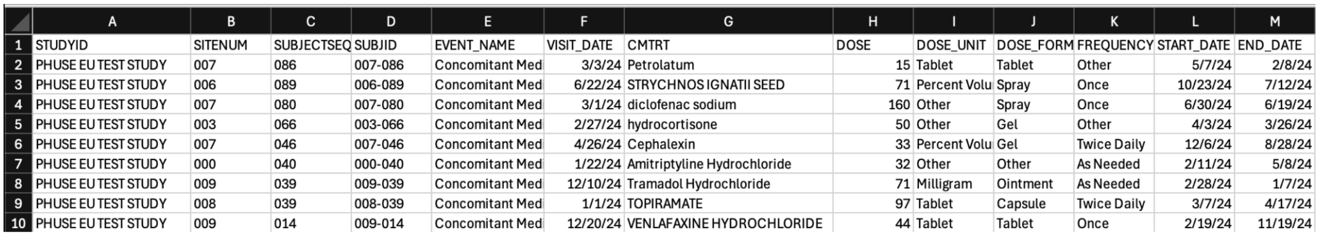
Use AI to generate fields based on the topic of your choice or example data.

Describe your topic: Examples: "flight logs", "social media", "stock trades"

Fields: 10

Or provide some example data: Paste CSV, JSON, or XML here...

Figure 4: Mockaroo's AI assisted configuration page



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	STUDYID	SITENUM	SUBJECTSEQ	SUBJID	EVENT_NAME	VISIT_DATE	CMTRT	DOSE	DOSE_UNIT	DOSE_FORM	FREQUENCY	START_DATE	END_DATE
2	PHUSE EU TEST STUDY	007	086	007-086	Concomitant Medi	3/3/24	Petrolatum	15	Tablet	Tablet	Other	5/7/24	2/8/24
3	PHUSE EU TEST STUDY	006	089	006-089	Concomitant Medi	6/22/24	STRYCHNOS IGNATII SEED	71	Percent Volu	Spray	Once	10/23/24	7/12/24
4	PHUSE EU TEST STUDY	007	080	007-080	Concomitant Medi	3/1/24	diclofenac sodium	160	Other	Spray	Once	6/30/24	6/19/24
5	PHUSE EU TEST STUDY	003	066	003-066	Concomitant Medi	2/27/24	hydrocortisone	50	Other	Gel	Other	4/3/24	3/26/24
6	PHUSE EU TEST STUDY	007	046	007-046	Concomitant Medi	4/26/24	Cephalexin	33	Percent Volu	Gel	Twice Daily	12/6/24	8/28/24
7	PHUSE EU TEST STUDY	000	040	000-040	Concomitant Medi	1/22/24	Amitriptyline Hydrochloride	32	Other	Other	As Needed	2/11/24	5/8/24
8	PHUSE EU TEST STUDY	009	039	009-039	Concomitant Medi	12/10/24	Tramadol Hydrochloride	71	Milligram	Ointment	As Needed	2/28/24	1/7/24
9	PHUSE EU TEST STUDY	008	039	008-039	Concomitant Medi	1/1/24	TOPIRAMATE	97	Tablet	Capsule	Twice Daily	3/7/24	4/17/24
10	PHUSE EU TEST STUDY	009	014	009-014	Concomitant Medi	12/20/24	VENLAFAXINE HYDROCHLORIDE	44	Tablet	Tablet	Once	2/19/24	11/19/24

Figure 5: Example csv output from the above configuration

STRENGTHS

1. **Cost-Effective**: Mockaroo offers a free version, making it accessible for various project sizes and budgets. The free version has enough functionality to get started using this tool and the paywalled features mostly impact scalability.
2. **Quickly Generates Lots of Data**: Once configured, Mockaroo can quickly generate large volumes of test data suitable for stress testing and comprehensive scenario coverage.
3. **Highly Configurable**: Users can provide just column names and data types, or offer more detailed specifications including custom lists and ranges for numeric values. It supports a wide range of data types, including drug names for realistic coding scenarios.
4. **AI-Assisted Configuration**: Mockaroo uses AI to guess data types and ranges from sample data, reducing setup time.

WEAKNESSES

1. **Time Investment in Configuration**: Initial setup can be time-consuming and there's a bit of a learning curve to the system. Luckily there are AI-powered solutions for accelerating this configuration, but unfortunately are most effective when there is already a few rows of test data in the EDC
2. **Requires Understanding of EDC Configuration:** This approach requires intimate knowledge of the EDC build to ensure data accuracy and relevance. Without direct integration with the EDC system, the method doesn't inherently reflect the EDC's gating logic, potentially allowing for impossible data combinations. The manual entry will not enforce the specific codelists embedded in the EDC, leading to potential inconsistencies (e.g., varying formats for sex: m/f, M/F, Male/Female, 1/0).
3. **Inflexibility to EDC Changes:** As the clinical trial evolves and the EDC structure is modified, the configuration may quickly become outdated, requiring manual updates.
4. **Specific Limitations**: Some data scenarios, such as partially missing dates, may not be supported, and there's a risk of generating nonsensical data in certain situations.

SUMMARY

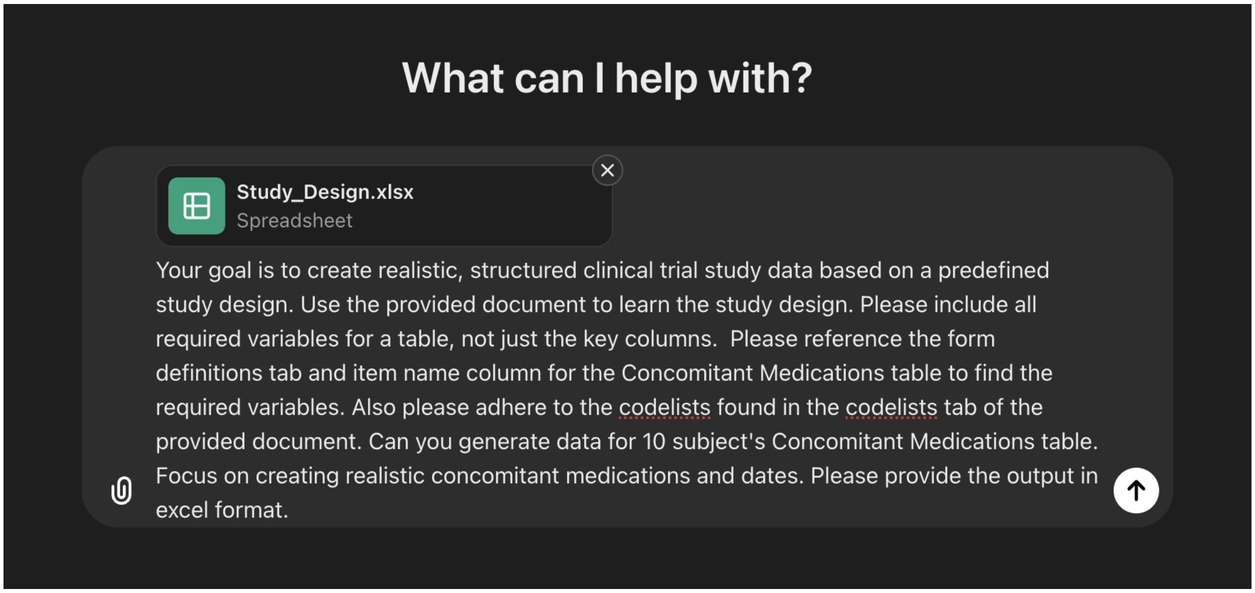
Mockaroo offers a powerful and flexible solution for generating large volumes of test data quickly, with AI-assisted configuration to ease the setup process. While it provides extensive customization options and supports a wide range of data types, it requires careful configuration to accurately reflect EDC structures and logic. The tool's efficiency in generating high volumes of data must be balanced against the need for manual updates when EDC changes occur and the potential for inconsistencies with EDC-specific features. Despite these limitations, Mockaroo represents a significant step towards more efficient test data generation, particularly for scenarios requiring large datasets or complex data types.

SOLUTION 4: LLM DATA GENERATION WITH CHATGPT

METHOD

This approach uses ChatGPT, a large language model (LLM), to generate realistic test data for clinical studies. The process begins with uploading a study design document in Excel format, which contains pre-specified forms and code lists, typically available as an export from any EDC system. Users then craft a prompt for ChatGPT, providing context, specific requirements, and references to the uploaded document. The AI model interprets this information to generate test data that aligns with the study design and specifications. Crafting an effective prompt is crucial for generating high-quality test data. A well-constructed prompt typically includes the following elements:

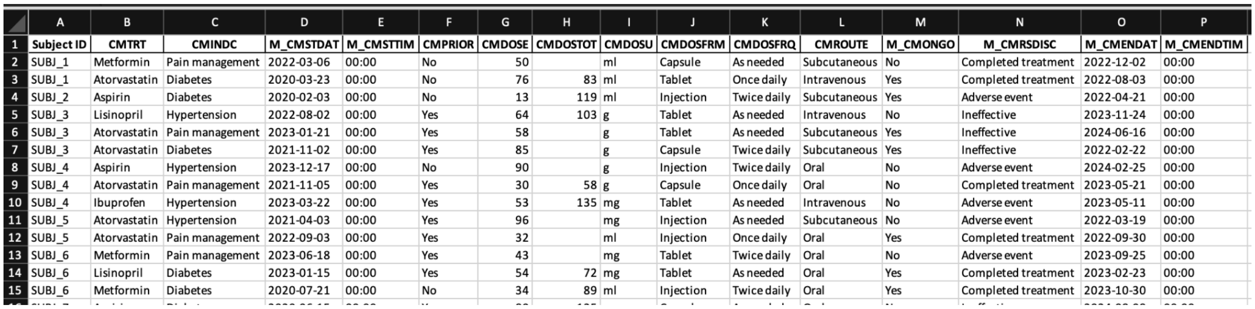
1. **Context and specificity**: Provide background about the clinical trial and clearly define required data elements, formats, and constraints.
2. **Data requirements**: Specify the volume of data needed, any required edge cases or scenarios, and emphasis on consistency with study-specific standards.
3. **Reference to uploaded document**: Explicitly instruct ChatGPT to use the information from the study design document, including adherence to codelists and data relationships.



What can I help with?

Your goal is to create realistic, structured clinical trial study data based on a predefined study design. Use the provided document to learn the study design. Please include all required variables for a table, not just the key columns. Please reference the form definitions tab and item name column for the Concomitant Medications table to find the required variables. Also please adhere to the codelists found in the codelists tab of the provided document. Can you generate data for 10 subject's Concomitant Medications table. Focus on creating realistic concomitant medications and dates. Please provide the output in excel format.

Figure 6: Prompt and study design document provided to ChatGPT in order to output test data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Subject ID	CMTRT	CMINDC	M_CMSTDAT	M_CMSTTIM	CMPRIOR	CMDOSE	CMDOSTOT	CMDOSU	CMDOSFRM	CMDOSFRQ	CMROUTE	M_CMONGO	M_CMRSDISC	M_CMENDAT	M_CMENDTIM
2	SUBJ_1	Metformin	Pain management	2022-03-06	00:00	No	50		ml	Capsule	As needed	Subcutaneous	No	Completed treatment	2022-12-02	00:00
3	SUBJ_1	Atorvastatin	Diabetes	2020-03-23	00:00	No	76	83	ml	Tablet	Once daily	Intravenous	Yes	Completed treatment	2022-08-03	00:00
4	SUBJ_2	Aspirin	Diabetes	2020-02-03	00:00	No	13	119	ml	Injection	Twice daily	Subcutaneous	Yes	Adverse event	2022-04-21	00:00
5	SUBJ_3	Lisinopril	Hypertension	2022-08-02	00:00	Yes	64	103	g	Tablet	As needed	Intravenous	No	Ineffective	2023-11-24	00:00
6	SUBJ_3	Atorvastatin	Pain management	2023-01-21	00:00	Yes	58		g	Tablet	As needed	Subcutaneous	Yes	Ineffective	2024-06-16	00:00
7	SUBJ_3	Atorvastatin	Diabetes	2021-11-02	00:00	Yes	85		g	Capsule	Twice daily	Subcutaneous	Yes	Ineffective	2022-02-22	00:00
8	SUBJ_4	Aspirin	Hypertension	2023-12-17	00:00	No	90		g	Injection	Twice daily	Oral	No	Adverse event	2024-02-25	00:00
9	SUBJ_4	Atorvastatin	Pain management	2021-11-05	00:00	Yes	30	58	g	Capsule	Once daily	Oral	No	Completed treatment	2023-05-21	00:00
10	SUBJ_4	Ibuprofen	Hypertension	2023-03-22	00:00	Yes	53	135	mg	Tablet	As needed	Intravenous	No	Adverse event	2023-05-11	00:00
11	SUBJ_5	Atorvastatin	Hypertension	2021-04-03	00:00	Yes	96		mg	Injection	As needed	Subcutaneous	No	Adverse event	2022-03-19	00:00
12	SUBJ_5	Atorvastatin	Pain management	2022-09-03	00:00	Yes	32		ml	Injection	Once daily	Oral	Yes	Completed treatment	2022-09-30	00:00
13	SUBJ_6	Metformin	Pain management	2023-06-18	00:00	Yes	43		mg	Tablet	Twice daily	Oral	No	Adverse event	2023-09-25	00:00
14	SUBJ_6	Lisinopril	Diabetes	2023-01-15	00:00	Yes	54	72	mg	Tablet	As needed	Oral	Yes	Completed treatment	2023-02-23	00:00
15	SUBJ_6	Metformin	Diabetes	2020-07-21	00:00	No	34	89	ml	Injection	Twice daily	Oral	Yes	Completed treatment	2023-10-30	00:00

Figure 7: Example excel output from ChatGPT showing all of the required CM fields and realistic data

STRENGTHS

1. **Rapid Data Generation:** ChatGPT can produce large volumes of test data in minutes, dramatically reducing the time required compared to manual methods or traditional automated tools. Additionally, the speed of generation allows for quick iterations and refinements, enabling rapid adjustments to meet changing study requirements and test data generation needs
2. **Quality and Consistency**: Because we provide the study design to ChatGPT, the generated data adheres to the given study structure, complying with specified codelists, date formats, and common fields. This results in higher quality data sooner in the process compared to manual methods.
3. **Adaptability**: The method easily accommodates changes in the EDC build, as it references the most current build specifications from the source system. In addition, well-crafted prompts allow for highly specific and contextualized data generation tailored to unique study requirements.
4. **Cost-Effective**: While there are paid enterprise versions of ChatGPT, we were able to achieve quality results using the free version of the tool

WEAKNESSES

1. **Learning Curve**: The quality of generated data is heavily reliant on the quality and specificity of the prompts used. Learning how to write these high quality prompts is a skill that takes time to learn. It may also take time to perfect prompts for specific data formats, requiring multiple iterations to achieve desired results.
2. **Context Limitations**: While generating realistic data, the AI might lack nuanced understanding of patient variability or study-specific intricacies. Without careful prompt design, there's a risk of generating data that doesn't fully align with complex study protocols or real-world clinical patterns.

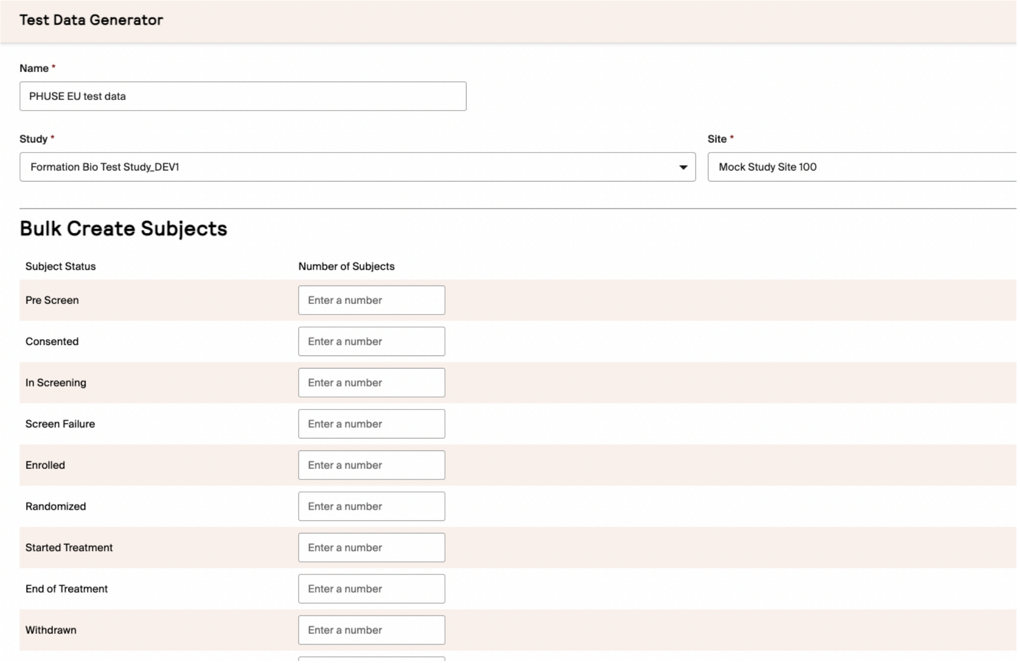
SUMMARY

ChatGPT-based data generation offers a rapid, cost-effective approach to creating large volumes of realistic and consistent test data for clinical trials, dramatically reducing time compared to manual methods. Its ability to quickly accommodate changes in study design and generate high-quality data that adheres to specifications provides a significant advantage, enabling fast iterations and refinements. However, the method's effectiveness heavily depends on the user's skill in crafting high-quality prompts, presenting a learning curve that requires time and practice to master. Additionally, while ChatGPT can generate data rapidly, it may struggle with capturing nuanced complexities of patient variability and intricate study-specific details without careful prompt design. Despite these challenges, the combination of speed, quality, adaptability, and cost-effectiveness positions ChatGPT as a valuable tool in the test data generation landscape.

SOLUTION 5: LLM GENERATED DATA INTEGRATED WITH EDC

METHOD

This approach uses custom software to integrate LLM capabilities with the EDC. At Formation Bio, we use Veeva as our EDC, which sports a robust API for data input and study spec export. Using these API integrations, we can pull the latest version of the study design document out of Veeva, allowing for up-to-date forms for all subjects. The system can then programmatically ask the LLM to generate data for each of the provided forms in their exact formats; since we can programmatically iterate over all forms for all visits, it means we can generate an entire subject's worth of data in a single operation - usually in a matter of minutes. Using LLMs, the system is able to prompt for various profiles of patients, crafting narratives that can then be derived into test data, such as early termination or screen fail patients, which the user can specify the quantities for the number of each type of subject they may want at the beginning of the run. Once the user reviews the generated data and makes any desired manual edits, the system then programmatically re-enters that data back into the EDC using Veeva's API once again, creating the new subjects at a specified site matching the exact data generated. In this way, not only does the test data exist in a raw spreadsheet format, but it also lives in the system where data managers and clinical analysts will access it from.



Test Data Generator

Name *: PHUSE EU test data

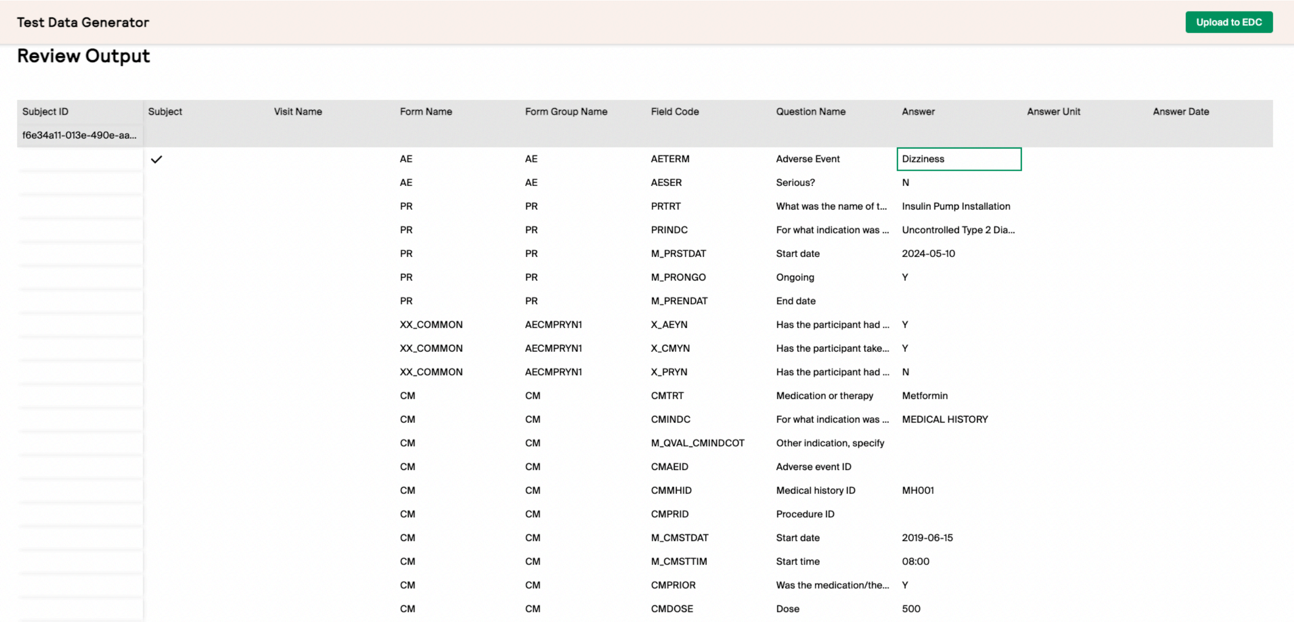
Study *: Formation Bio Test Study_DEV1

Site *: Mock Study Site 100

Bulk Create Subjects

Subject Status	Number of Subjects
Pre Screen	Enter a number
Consented	Enter a number
In Screening	Enter a number
Screen Failure	Enter a number
Enrolled	Enter a number
Randomized	Enter a number
Started Treatment	Enter a number
End of Treatment	Enter a number
Withdrawn	Enter a number

Figure 8: Screenshot of generate run screen where the user can select the number of subjects per status to generate data for



Test Data Generator

Review Output

Subject ID	Subject	Visit Name	Form Name	Form Group Name	Field Code	Question Name	Answer	Answer Unit	Answer Date
f6e34a11-013e-490e-aa...									
	✓		AE	AE	AETERM	Adverse Event	Dizziness		
			AE	AE	AESER	Serious?	N		
			PR	PR	PRTRT	What was the name of t...	Insulin Pump Installation		
			PR	PR	PRINDC	For what indication was ...	Uncontrolled Type 2 Dia...		
			PR	PR	M_PRSTDAT	Start date	2024-05-10		
			PR	PR	M_PRONGO	Ongoing	Y		
			PR	PR	M_PRENDAT	End date			
			XX_COMMON	AECMPRYN1	X_AEYN	Has the participant had ...	Y		
			XX_COMMON	AECMPRYN1	X_CMYN	Has the participant take...	Y		
			XX_COMMON	AECMPRYN1	X_PRYN	Has the participant had ...	N		
			CM	CM	CMTRT	Medication or therapy	Metformin		
			CM	CM	CMINDC	For what indication was ...	MEDICAL HISTORY		
			CM	CM	M_QVAL_CMINDCOT	Other indication, specify			
			CM	CM	CMAEID	Adverse event ID			
			CM	CM	CMMHID	Medical history ID	MH001		
			CM	CM	CMPRID	Procedure ID			
			CM	CM	M_CMSTDAT	Start date	2019-06-15		
			CM	CM	M_CMSTTIM	Start time	08:00		
			CM	CM	CMPRIOR	Was the medication/the...	Y		
			CM	CM	CMDOSE	Dose	500		

Figure 9: Screenshot of data review screen where the user can review the data generated and make any updates to the “Answer” column. This is an optional step before uploading directly to the EDC

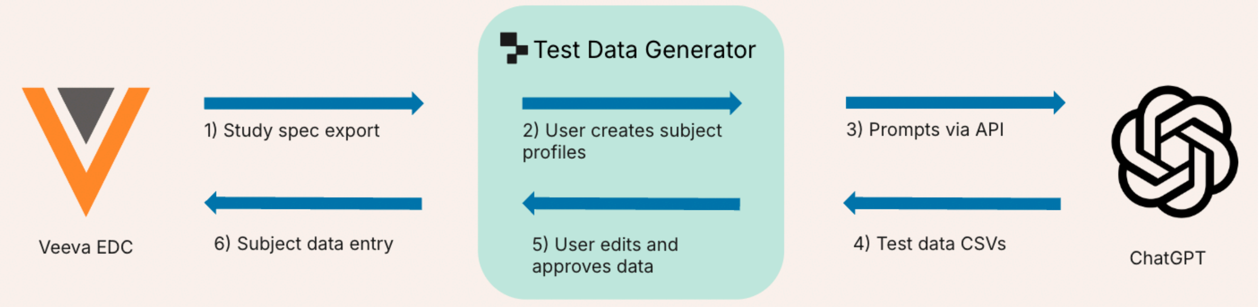


Figure 10: Overview of the design of the system

STRENGTHS

- **Immense Speed and Scale**: This integrated system can create a full subject’s study dataset in a matter of minutes, and can create multiple subjects in a single operation and with limited human intervention.
- **EDC Connectivity**: By tying directly to an ongoing study build, it is assured that test data is always generated off of the latest forms and rules, allowing for painless data regeneration through study start-up. With its ability to write data back into the EDC as well, test data can also be duplicated back into the system where it will ultimately be entered during a study, assisting with UAT efforts and enabling test runs of downstream capabilities.
- **User Experience**: With a simple web application, a clinical user is capable of specifying the number of subjects for each status as well as which site the subjects should be generated in. After the data has been generated, the user may additionally alter the data to add additional edge cases and errors before pushing that data to the EDC system.
- **Unlocked Capabilities**: With a way to create multiple full subjects of test data quickly, data analysts and study teams are capable of performing tasks far earlier in a study than before. Teams can create enough data to begin statistical programming, clinical monitoring, and risk-based quality management efforts prior to live subject data being collected. It also opens up the ability to run a mock study with test data, trialing the standard operating procedures of a study team during study build.

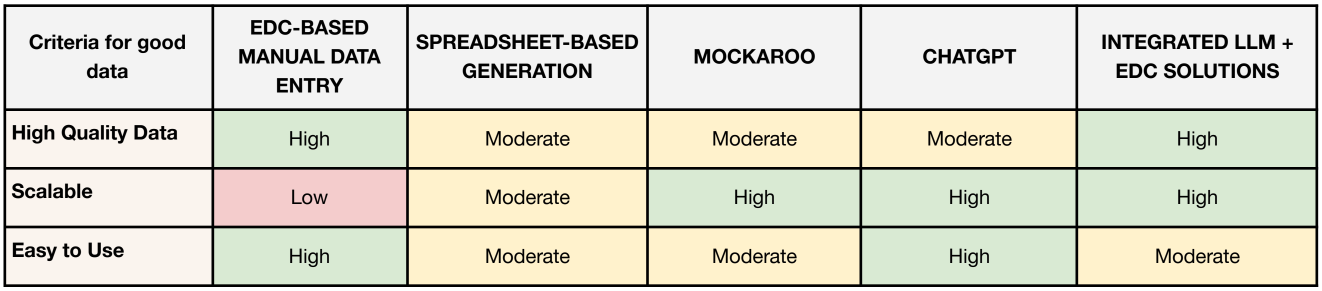
WEAKNESSES

- **Limited EDC Choices:** Due to the sophistication of APIs needed to support these types of programs, only certain EDC systems can be utilized here.
- **Engineering Requirements**: Custom code must be written to connect the LLM and EDC technologies, as well as process the data and enable humans to interact. Thus, trained engineers are required to develop and support these systems.

SUMMARY

A fully integrated software solution offers the most in terms of speed and scale of test data creation. By programmatically connecting directly to a live study build, both its specifications and its data entry, we ensure accuracy and precision in the data matching the live build, overcoming much of the cost of having to remake test data when the build changes. By putting data directly into the live system, it also enables a holistic quality assurance process to not only confirm the validity of the study build, but also to enable downstream development and testing of scripts, programs, monitoring tools, and standard operating procedures put in place by various teams during a study. By shifting left on this process, allowing for development and testing of study monitoring activities sooner in study start-up, it creates the ability to have a high quality in place by Day 1, ensuring the correct patient safety measures are in place for even the first patients. However, this approach requires significant investments from a study team, both in technically-adept EDC vendors and engineering resources to develop such an integrated tool. Nevertheless, overcoming these costs presents a potential game changer in data quality and innovation for a study team.

CONCLUSION



Criteria for good data	EDC-BASED MANUAL DATA ENTRY	SPREADSHEET-BASED GENERATION	MOCKAROO	CHATGPT	INTEGRATED LLM + EDC SOLUTIONS
High Quality Data	High	Moderate	Moderate	Moderate	High
Scalable	Low	Moderate	High	High	High
Easy to Use	High	Moderate	Moderate	High	Moderate

Figure 11: Summary of the comparative analysis of test data generation options

Despite the growing buzz around AI, its practical application in clinical trials remains limited, with many processes still relying on traditional manual approaches. This paper presents test data generation as a compelling entry point for AI adoption in clinical trials, demonstrating how AI tools can transform a time-intensive manual process into one that takes minutes while simultaneously improving quality. This dramatic improvement in speed and quality scales with technology rather than headcount, offering a sustainable path forward for clinical trial operations. The progression from basic spreadsheet-based approaches to sophisticated LLM-integrated systems reveals that organizations can achieve both quality and speed through strategic investment in AI technologies, though this comes with an increasing learning curve.

The investment in AI capabilities, including the initial learning curve and engineering resources, yields substantial returns beyond mere operational efficiency. By empowering Clinical Data Analysts through targeted training programs, hackathons, and dedicated innovation time, organizations can build a culture of continuous improvement and technological advancement. When teams spend less time on manual data generation, they can focus on innovation and process improvement, leading to more robust code, comprehensive testing, and higher quality deliverables. This shift from manual effort to strategic thinking positions organizations to deliver more efficient, reliable, and innovative clinical trials. As the industry begins its AI journey, test data generation represents an accessible starting point with immediate, tangible benefits – providing a practical blueprint for broader AI adoption in clinical trial operations.

REFERENCES

1. Robust Test Data (https://phuse.s3.eu-central-1.amazonaws.com/Archive/2018/Connect/EU/Frankfurt/PAP_PP11.pdf)

RECOMMENDED READING

1. Mockaroo (https://www.mockaroo.com/)
2. ChatGPT (https://chatgpt.com/)
3. OpenAI APIs (https://platform.openai.com/docs/guides/text-generation)
4. Veeva CDMS APIs (https://developer-cdms.veevavault.com/api/24.2/#getting-started)

CONTACT INFORMATION

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

Author Name: Emily Yates
Company: Formation Bio
Email: eyates@formation.bio
Website: https://www.formation.bio/

Author Name: Matt Luppino
Company: Formation Bio

Email: mluppino@formation.bio
Website: https://www.formation.bio/

Author Name: Andrew Burd
Company: Formation Bio
Email: aburd@formation.bio
Website: https://www.formation.bio/