Paper ML25

Accelerating Study Build Using AI Enabled CRF Design

Aimee Basile, Otsuka Pharmaceuticals, Princeton, NJ, USA
Jeremy Zhang, Otsuka Pharmaceuticals, Princeton, NJ, USA

ABSTRACT

Industry benchmarks show that, on average, it takes 4 months from protocol approval to first patient, first visit for Phase 1 studies. For Phase 1 studies it is critical to reduce this front-end time to speed up the clinical development of drug products. The industry has been looking at ways to reduce this time with digitization and standardization of the protocol which helps decrease the time for protocol development and supports study design and build. To further decrease study build timelines we are piloting a GenAI solution to digest the protocol and Schedule of Activities (SoA) and create draft CRFs based on a standard library of data collection metadata, historical CRF used in previous studies and de novo generated CRFs. This first draft of CRFs from the SoA allows parallel development of CRFs and protocol and should decrease the study team review time for a savings of up to four weeks, reducing the timeline to three months.

1.0 INTRODUCTION

1.1 PROBLEM STATEMENT

With industry focus on early phase and rapid development, the goal is to reduce early phase study timelines. The longest time for an early phase study is the study database build, which can take the same amount of time as for later phase studies. The root cause of that delay is that the CRF development process is still very manual with many iterations of CRF review. While implementing standard CRFs can streamline the process, there are still enough variations in CRFs that require time to develop and extend the desired timeline.

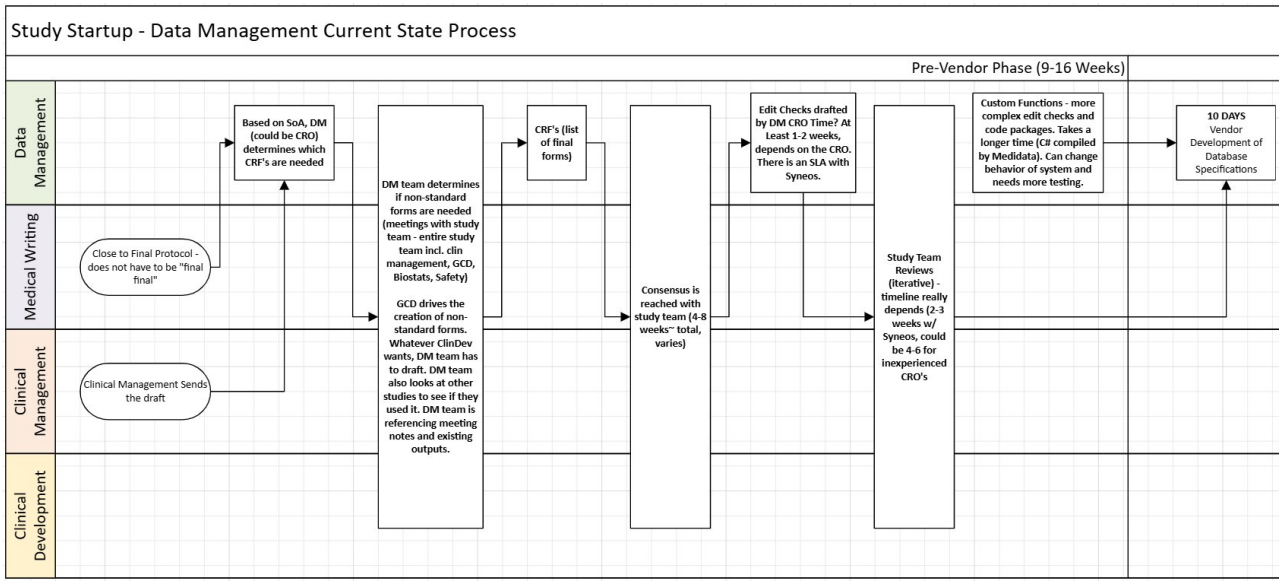
1.2 SOLUTION

Industry standard organizations have been working on solutions for streamlining protocol to CRF development by using USDM and JSON to create a digitized version of the SoA; however, there has been limited uptake by Medical Writing groups to use those methods. Therefore, we have decided to utilize an AI solution to digitize a draft version of the protocol to generate draft CRFs. This GenAI plugin for Microsoft Office utilizes current standard CRFs, historical CRFs and de novo generated CRFs. Once the CRFs are reviewed, the tool will also generate edit check specifications for those CRFs, and we expect a reduction of overall study conduct timelines by 2-4 weeks.

2.0 PRIOR PROCESS VS. FUTURE PROCESS

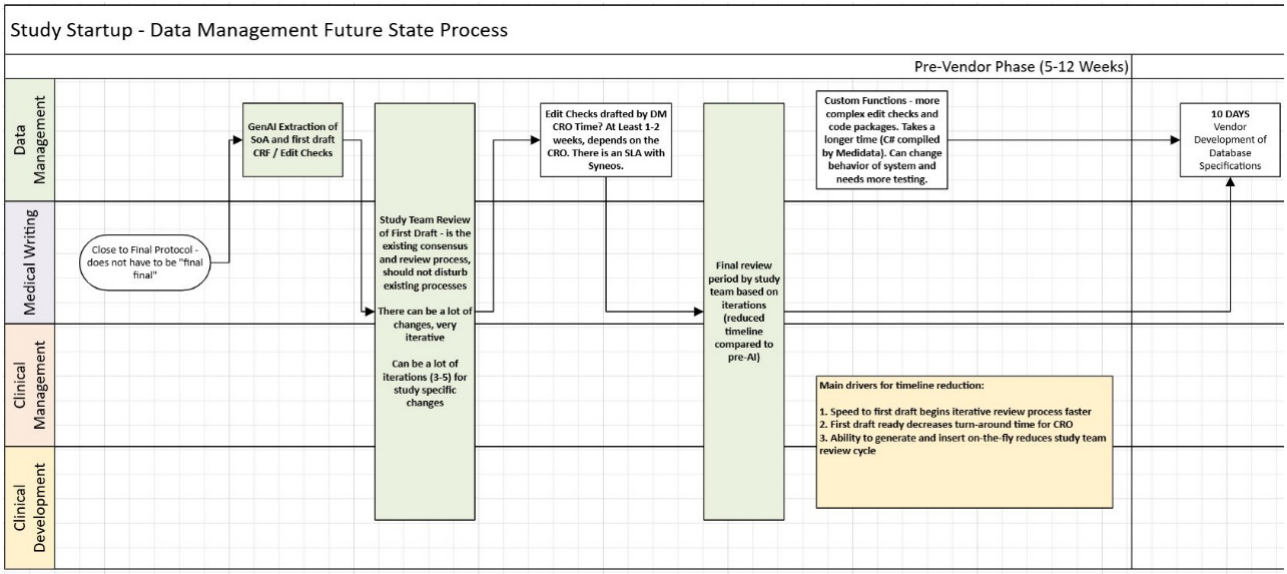
The prior process was to use a draft version of the protocol to begin the study database design process; however, the draft version was close to final and simply avoided the approval process to save a few days. In general, the protocol was sent to a CRO partner to begin the development of the study database using their standard CRFs plus any historical CRFs from previous studies for the compound. Any new CRFs would need to be developed from scratch with the clinical study team. Once CRFs were reviewed and approved, edit checks would be generated while the database was set up. Interactive review was done on the CRFs to ensure the study team approved of the CRFs. Finally, database would be completed with the finalized CRFs, edit checks and any custom programming needed. This process typically took about five months.

FIGURE 1. CURRENT PROCESS FOR STUDY INITIATION



The new process will start with the protocol with only table of contents and a schedule of assessments required to generate the CRFs. CRFs can be generated using the AI tool in about 30 minutes. The study team will still need to review the CRFs and approve them, but the AI generation of de novo CRFs will reduce the amount of time for creating new CRFs by serving as a foundation for the final design. In addition, once the CRFs are approved, they are used to feed into another AI tool to generate the appropriate edit checks for the study. We expect these process changes and technical tools will combine to reduce the study database build timelines by at least 4 weeks.

FIGURE 2. FUTURE PROCESS FOR STUDY INITIATION

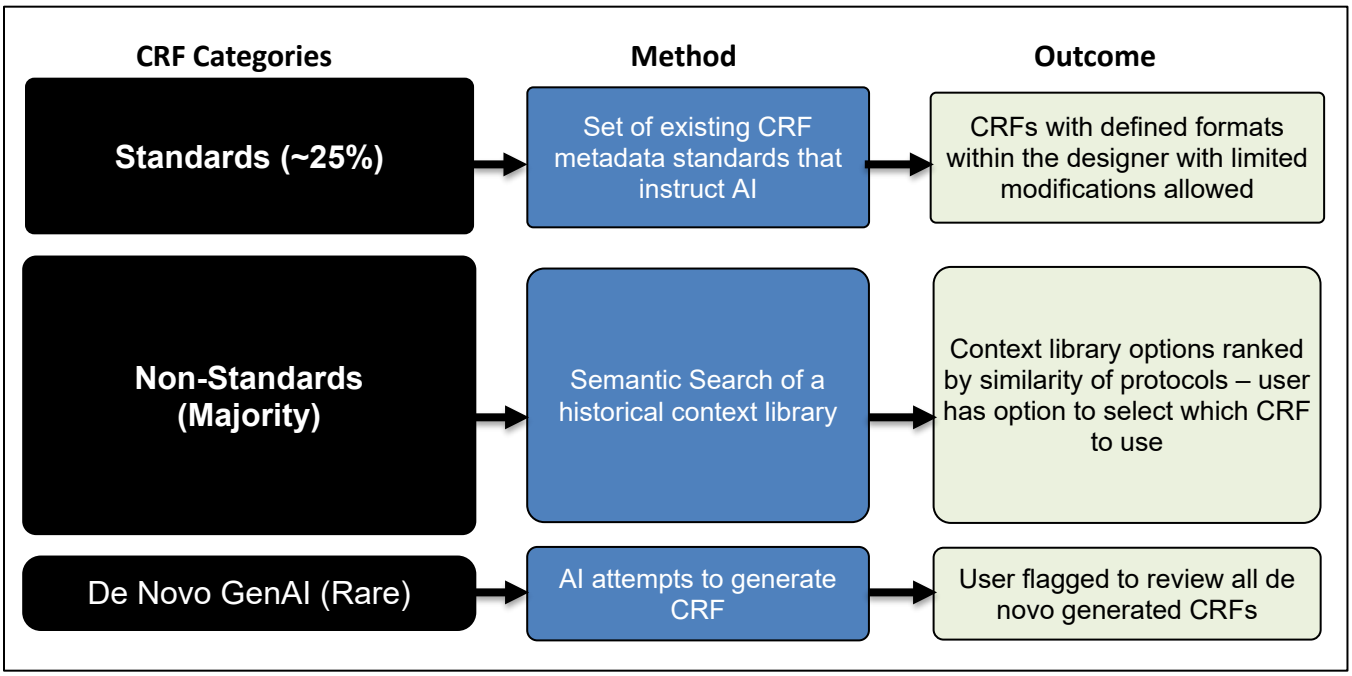


3.0 HOW STANDARDS ARE UTILIZED IN THE PROCES

With the use of standard CRF metadata, the AI generator tool prioritizes matching the standards when one of the standard domains is referenced in the protocol. In general, about 25% of the CRFs generated come from standard CRF metadata. As additional standard CRF metadata are generated, we expect the matches with the standards to increase by up to 80%. The more we standardize, the better the AI can get at generating study CRFs.

If standard metadata is not matched with the protocol, a semantic search of a historical context library is used and the majority of the AI generated CRFs currently fall into this category. Finally, if a match cannot be found in the standard or historical catalogs, then the AI solution will generate a new CRF. The user is flagged to review all de novo generated CRFs, and all CRFs can be updated for study-specific details within the Microsoft Word AI Writing Tool.

FIGURE 3. USE OF STANDARDS IN THE AI TOOL

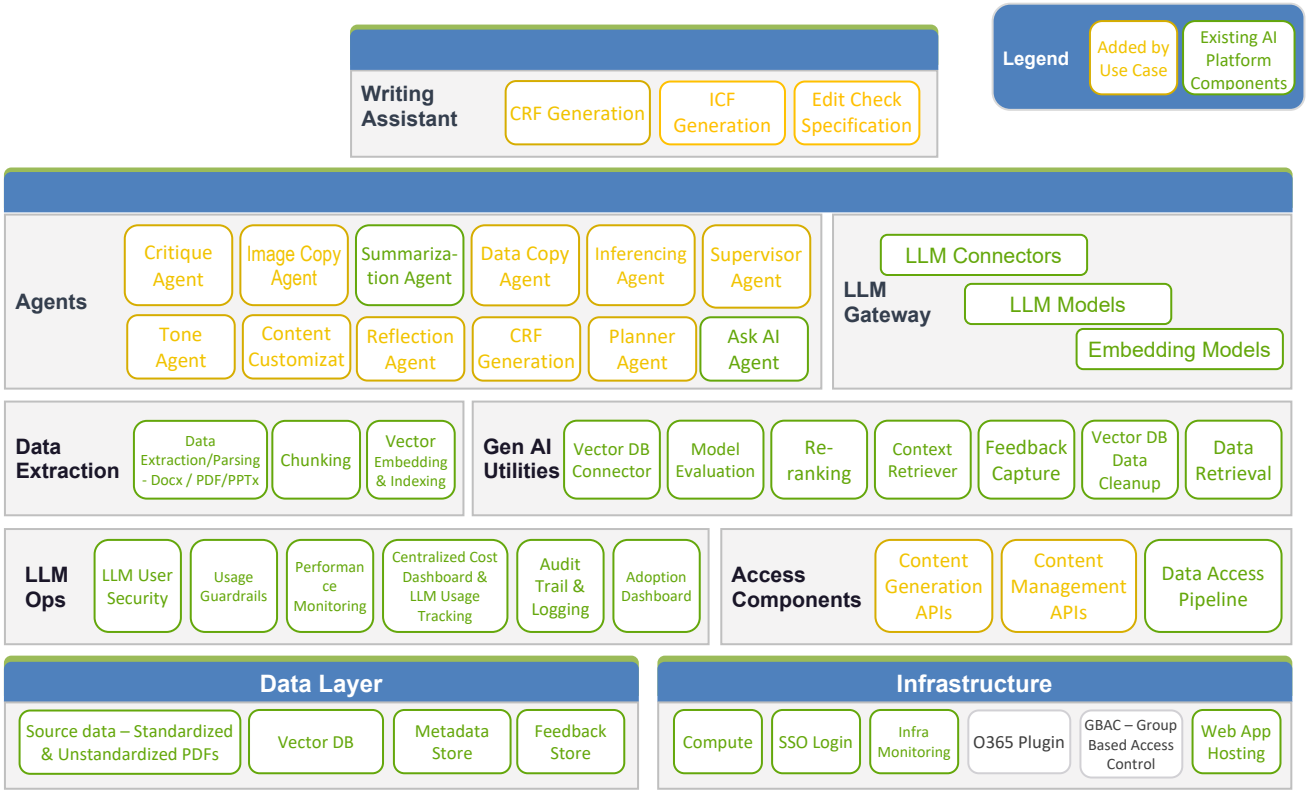


4.0 ARCHITECURAL AND AI TECHNOLOGY

4.1 SOLUTION ARCHITECTURE

The architecture of the AI solution is based on an enterprise AI platform and layered on top of a foundational data layer, infrastructure layer, and AI platform utilities. Across AI use cases, common reusable capabilities such as chunking, vector databases, and LLM connectors are customized and deployed as needed for individual solutions based on the solution design fit for purpose for each use case. Use cases may also contribute additional capabilities to the enterprise AI platform such as specific agents that have narrow scope or access to limited tools.

FIGURE 4. AI AGENTS



The CRF generation solution contributed several novel capabilities to the enterprise platform, including the concept of a planner agent for content generation, reusable assets for protocol digitization, and agents that enable content customization and critique. These agents have since been used in other capabilities such as medical and clinical document generation.

4.2 AGENTS AND FUNCTIONS

The AI platform utilizes an open-source agent orchestration framework based on LangGraph with the concept of agents acting as nodes with access to tools. Orchestration begins with a planner agent that utilizes specific user inputs to determine the target document and source document context necessary for generation. The generation of individual CRFs behaves like a source to target mapping. The source document – clinical protocol – is parsed and the schedule of assessments as well as inclusion and exclusion criteria are digitized. The planner delegates the identification of CRFs to generate to the data retrieval agent which uses a combination of AI-based document extraction, rules, and machine learning to identify procedures from both the schedule of assessments table as well as footnotes. For each procedure identified, they are standardized against known procedures from an existing context library of past CRFs. Agents such as the data retrieval agent have access to specific tools and utilities like SOA extraction utilities, LLM function utilities, OpenSearch embeddings, text chunkers, text digitizers, and database query utilities.

Semantic matching of vectorized inputs against context libraries in the OpenSearch vector databases is used across agents for tasks such as discovering the most appropriate standard procedure from the SoA table, matching the most relevant standard to a procedure, and matching the most relevant CRF and fields for a particular procedure. The matching algorithm ranks potential matches, and the matching scores can be made available to users.

4.3 CONTEXT LIBRARY

The context library contains different elements organized inside OpenSearch vector databases. These are protocol summaries used for semantic matching with best historical CRFs and historical forms, fields, and edit checks used as references. For each protocol, critical metadata elements are also separately indexed such as study ID, indication, phase, therapeutic area, and patient population. There is also a set of standard procedure names used as reference to increase the probability of matching extracted procedures to standard and historical context forms.

4.4 AUTOMATED TESTING

Errors in an agentic AI system propagate – meaning even if some steps of the agentic AI process have high accuracy (>90%), the system as a whole may still be inaccurate and require significant manual review.

For CRF generation, a series of agentic AI steps were tested via automation scripts for accuracy. Two steps in the extraction of the schedule of assessments – both the frequency and accuracy of standard procedure terminology matching were assessed at greater than 95% each. For CRF mapping, the semantic mapping of extracted procedures to historical or standard CRFs, the formatting and completeness of CRFs generated, and the subjective quality assessed on a 5-point scale were all assessed at greater than 90% accuracy or quality.

In order to ensure that error propagation in the system does not decrease the quality of generated outputs below a certain threshold, a system-wide metric of less than 15% of CRFs requiring human-in-the-loop intervention (changes to forms or changes at a field level) was set as a system-wide requirement. User intervention is defined as any addition, removal, edit, or change to the first draft of the CRFs generated by the system at either a form or field level. When tested against 5 gold standard protocol CRFs, the overall system performance exceeded 95% accuracy with the first step (generation of correct procedures and CRFs) excluding field-level accuracy exceeding 99% accuracy.

FIGURE 5. AUTOMATED TESTING RESULTS

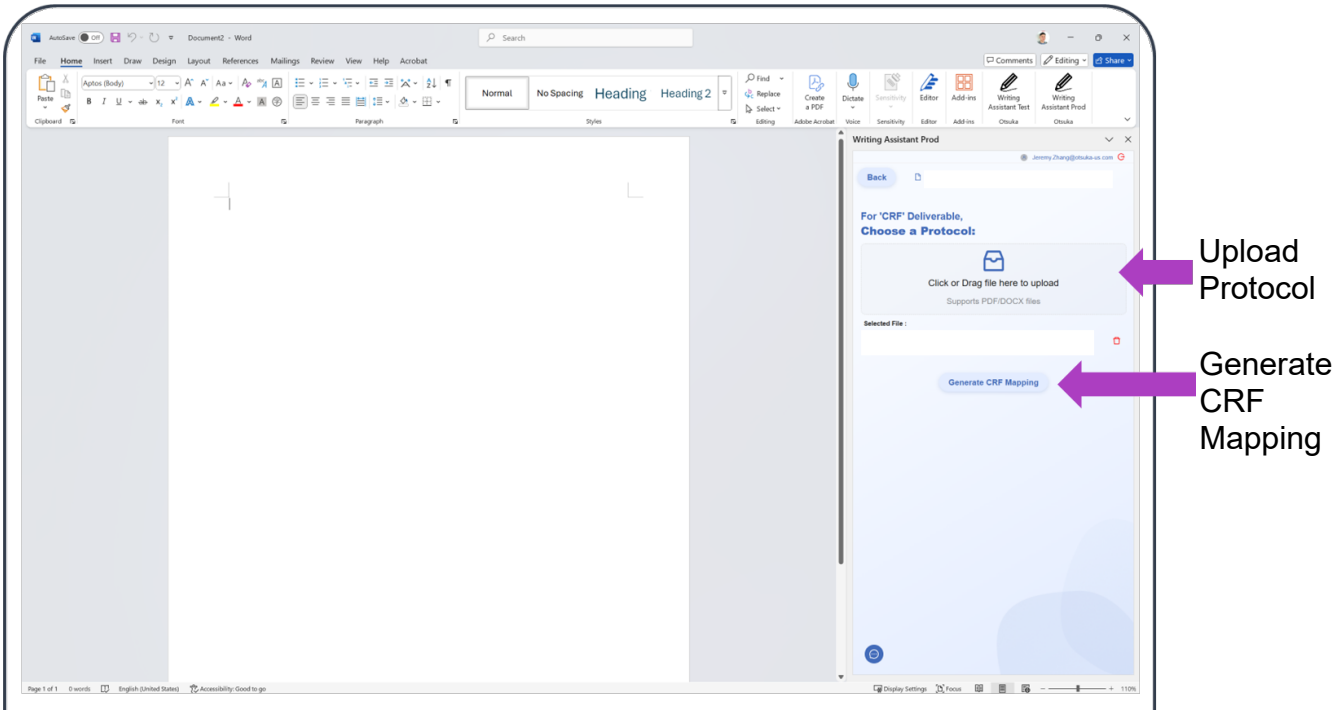
Testing Parameters			Gold Standard Testing Dataset (CRF)				Actual Output (CRF Tool)					Triage	
Protocol ID / Name	CRF Document Name	DateTime of Test	Expected Form Name	Expected Field Name	Expected Field Value	Expected CRF Type (Standard, Historical, LLM)	Output Form Name	Output Field Name	Output Field Value	Output CRF Type (Standard, Historical, LLM)	Comments / Attachments	Product Owner Triage	Development Team Comment
Protocol ID redacted		10:13 am|9/29	only 116/121 forms are generated but the bar shows 121										
		10:13 am|9/29	Adverse Events	forms are generated as expected	values are displaying as expected	Standard	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	standard			
		10:13 am|9/29	Prior and concomitant medications	forms are generated as expected	values are displaying as expected	standard	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	standard			
		10:13 am|9/29	demographics	forms are generated as expected	values are displaying as expected	standard	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	standard			
		10:13 am|9/29	Informed Consent	forms are generated as expected	values are displaying as expected	standard	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	standard			
		10:13 am|9/29	medical history	forms are generated as expected	values are displaying as expected	standard	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	standard			
		10:13 am|9/29	12-lead ECG	forms are generated as expected	values are displaying as expected	Historical	all the forms are generated as expected	fields are also generated as per standard CRF	values are also generated as expected	historical			
		10:13 am|9/29	IMP Administration			Historical	all the extracted forms	Randomization fields are present in addition to required fields	Randomization values are present in addition to required values	historical(384-201-00002_Annotated Unique CRF_04Nov2024.pdf)	attachment in sheet 6		
		10:13 am|9/29	Inclusion/exclusion			historical	screening -28to-2 form	footer values are dispalying as fields	page numbers from historical CRF is displaying in values	historical(384-201-00002_Annotated Unique CRF_04Nov2024.pdf)	attachment in sheet 6		

5.0 THE USER JOURNEY

5.1 CRF GENERATION

First, the user opens Microsoft Word and clicks on the Writing Assistant in the menu bar. Then, the user uploads a draft protocol and generates the CRFs.

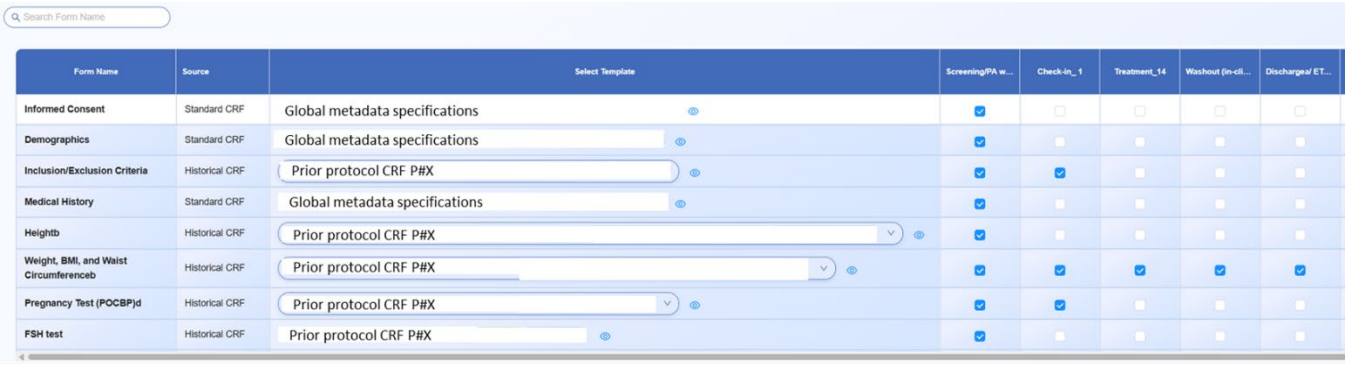
FIGURE 6: STARTING THE CRF GENERATOR



Once the CRFs are generated, a list of the CRFs selected will be presented. For historical CRFs the user can select the previous study they want to use for the historical CRF. This allows the user to select a previous study for the compound or indication that may be a more appropriate match for the CRF. Note that standard and de novo CRFs do not have template that can be selected. Also, CRFs can be added or deleted as needed from the generated CRFs.

In addition, the generated CRFs show which visits the CRF will appear in per the schedule of assessments. The user can select or deselect the visits for that CRF based on the protocol.

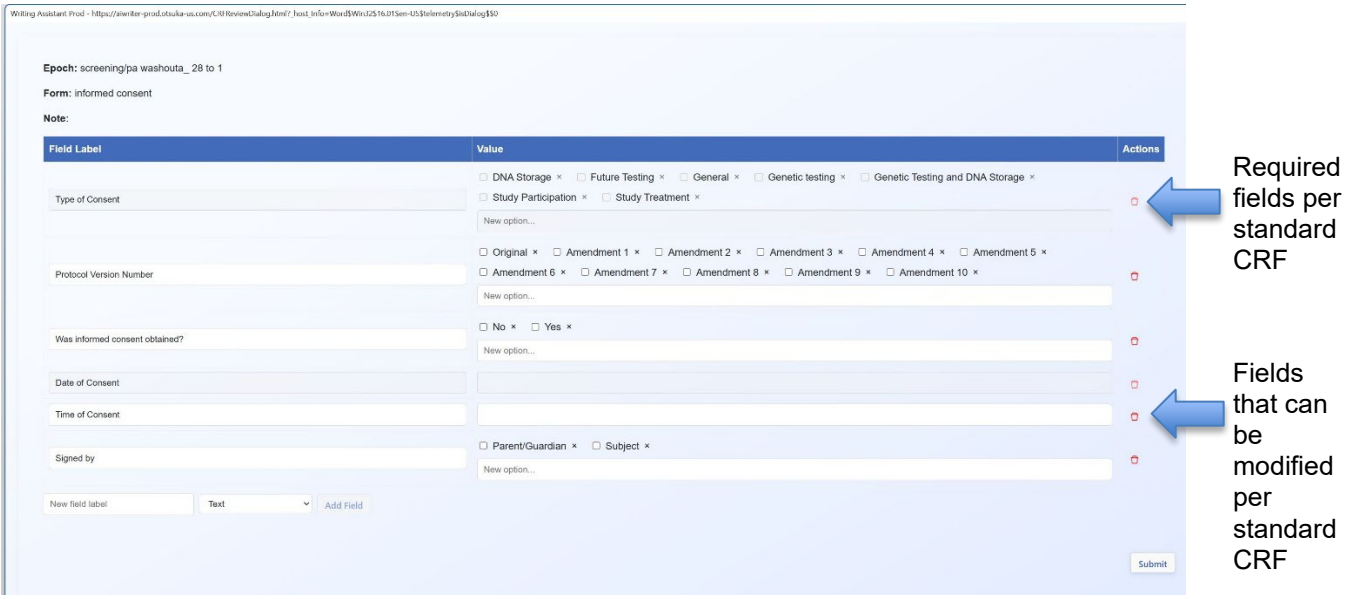
FIGURE 7. MAPPING RESULTS WITH VISIT SCHEDULE



Once the proper CRFs are selected for the correct visits, the user can then edit each CRF for the study. Standard CRFs have limited ability to make changes so that only allowable modifications can be made. For example, if a field is required to be on the CRF, per the standards, then it cannot be deleted or modified. If a field is optional or conditional it can be deleted or modified per the protocol.

For historical or de novo CRFs, the Field Label and the Values expected for that field can be changed. In addition, fields can be deleted and added, as needed.

FIGURE 8. EDITING A STANDARD CRF

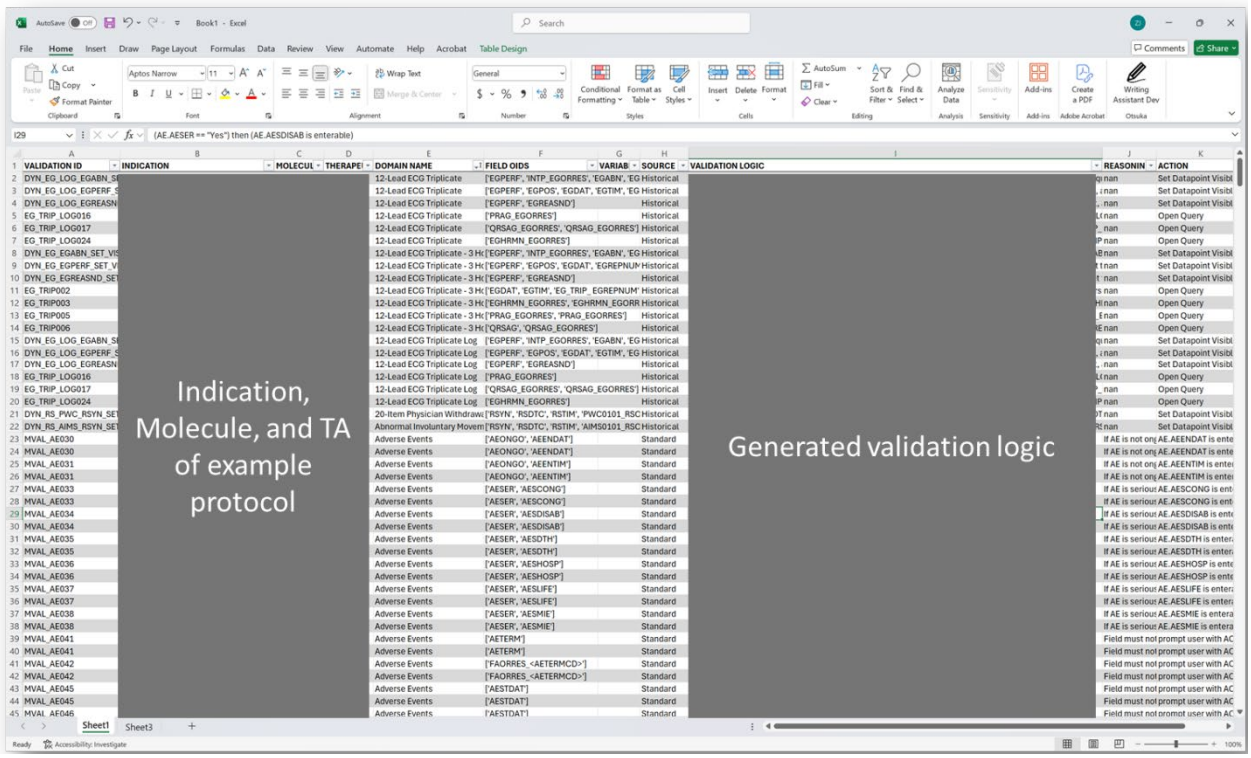


Once a CRF is edited, the user adds it to the Microsoft Word document for CRF review. At that time, the user can continue to refine the CRFs manually within Microsoft Word or can modify or further review the CRFs using common chat prompts in the AI chat engine.

5.2 EDIT CHECK GENERATION

A similar process within Microsoft Excel is used for generating the edit checks once the CRFs are finalized. The CRFs can be imported and digitized and associated edit checks are generated based on the CRFs and fields that have been approved by the clinical trial team. The AI solution uses the set of standard edit checks that are part of the standard CRFs and then references historical study database designs to match the study CRFs to previously defined edit checks for the same or similar data. Finally, the output of the AI generated edit checks can be reviewed and modified in Microsoft Excel using all of the features of Excel, such as filtering, sorting and even formula writing, if needed.

FIGURE 9. OUTPUT OF THE AI GENERATED EDIT CHECKS



CONCLUSION

This solution will allow the study database design process to start sooner in the study initiation phase and shorten the CRF and associated edit check design and review process. This solution also allows AI agents that already existed to be used for other purposes and build on those AI capabilities to eliminate the need for protocol digitization at the time of writing and eliminates the need for the protocol to be finalized before CRF design begins. In addition, this solution allows standard data collection metadata to be implemented in a user-friendly way without the overhead of a metadata repository system. In conclusion, this solution is lightweight and flexible, and is expected to reduce the study database build timeline by at least 4 weeks. The new process and AI solution are currently being piloted but will need more testing to know the actual time savings. The primary key to the success of this solution will be how change management implemented to ensure complete uptake and realize the maximum time savings.

CONTACT INFORMATION

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

Author Name: Aimee Basile
Company: Otsuka Pharmaceuticals, Inc.
Address: 508 Carnegie Center Dr, Princeton, NJ 08540
Email: aimee.basile@otsuka-us.com

Author Name: Jeremy Zhang
Company: Otsuka Pharmaceuticals, Inc.
Address: 508 Carnegie Center, Princeton, NJ 08540
Email: jeremy.zhang@otsuka-us.com