Paper ML03
Virtual Assistant for Conversational Experience with Clinical Data

Srinivasan Anandakumar, Saama Technologies, Bridewater (NJ), USA
Amit Gulwadi, Saama Technologies, Bridewater (NJ), USA

ABSTRACT

Clinical and scientific data collected from Electronic Data Capture (EDC) systems and third-party sources (lab data, biomarker and PK/PD data) during clinical trials involves intensive data exploration, visualization and analysis for ensuring data quality and patient safety. Current approaches typically utilize visualization tools where end users must navigate through multiple reports and dashboards to get the necessary insights. These reports and dashboards involve significant effort in building and integrating for every study, since they are typically integrated with study-specific data containers. An AI-based virtual assistant trained on both natural language and clinical trial standards (such as SDTM) enables the next generation conversational experience with clinical trial data. The virtual assistant platform that supports near real-time data standardization from EDC and third-party data into CDISC SDTM helps users with faster and more accurate insights (study specific and cross trial). Since the virtual assistant is pre-trained on SDTM standards, it provides an interactive experience for exploring and analyzing clinical trial data without a study specific integration and configuration effort. In this paper we will discuss the process, development approach and provide a case study of a clinical data virtual assistant proof-of-concept implementation.

INTRODUCTION

Typically the process involved in enabling clinical trials involving patient data and submissions is characterized with siloed, manual and repetitive tasks. Collected data variances across trials and metadata changes during trial conduct causes challenges to create a standardized process for enabling clinical trial analytics and a submission pathway. Clinical trial analytics involve analysis of a variety of disparate clinical trial source data, such as that from EDC and third-party sources (lab data, biomarker, and PK/PD data). Clinical trial analytic data sources are getting extended to newer types of data sources, such as OMICs, wearables and sensors, and un-structured data. Providing clinical trial analytics with an increase in scale and volume of data types is a challenge with the current study-specific reports and dashboards approach. Technologies enabled by machine learning offer approaches to near real-time standardized clinical pipelines and open up new communication channels (e.g. virtual assistant) for performing data analysis.

This paper is divided in 3 parts:

- Components of a next generation clinical data analytics platform (virtual assistant is a component of such a platform)
- Framework to build a clinical data virtual assistant
- Case study of a clinical data virtual assistant proof of concept

Components of a Next Generation Clinical Data Analytics Platform

For providing meaningful and timely insights, a virtual assistant needs to have near real-time standardized data. Standardized clinical data (instead of raw data) helps a virtual assistant to perform as

a plug-in module to provide conversational experiences on a new study. In this case, the virtual assistant can work on any SDTM study data without manual study set up tasks. But, near real-time generation of standardized clinical data has to account for complexities related to disparate study metadata and flux of study metadata due to protocol amendments. Generation of near real-time standardized data is a core feature of a next generation clinical data analytics platform. This section outlines the components of such a platform and outlines the component's role in creating near real –time standardized clinical data for virtual assistant.

Smart Data Ingestion – Smart data ingestion involves automated data acquisition from variety of legacy clinical trial sources, such as EDC, third party (lab and biomarker data) and newer data sources such as OMICS, wearables/sensors and semi and unstructured data sources. This is enabled by pre-built adapters from vendors, build on CDISC standards (e.g. ODM adapters for ingesting clinical data). The integration adapters should support a variety of data formats (e.g. XML, Parquet, flat files, SAS) and should support file-based and web services-based integration. The key feature of smart data ingestion is to automatically adapt to the changes in source metadata. This is enabled through a combination of metadata management and data quality modules. Key features of smart data ingestion include:

- Set up of structural and vocabulary metadata for the source files
- Detection of files for constructive and destructive change and auto-update of patient data lake based on metadata approval
- Auto-set up of study integration configuration based on global integration constructs
- Flexibility of extending integration with OMICS, unstructured data, documents and images, wearables and sensors through configuration

Metadata Management – Clinical data is associated with multiple types of metadata. They can be logically divided in to following types:

- Data Element Metadata – Precise definition of a data element across the organization
- Contextual Metadata – Usage of data element metadata to create logical structures for data collection, data submission and data analysis
- Transformation Metadata - Creation of hierarchical global maps (transformation maps cross global/therapeutic area/indication/project) for defining transformations for submission and analysis datasets

Apart from managing all the above three types of metadata, key features for metadata management include integration with industry sources of clinical metadata, such as CDISC SHARE, and management of contextual metadata across multiple levels of study hierarchy. The managed contextual metadata helps users to set up standards and create study-specific data models for storing standardized data.

Intelligent Mapping and Continuous Standardization – Study-specific programming for creating standardized clinical datasets (the current typical way) creates obstacles for creating a continuous near real-time clinical standardization pipeline. Using machine learning-based techniques to create study maps based on hierarchical global maps and study metadata can enable continuous generation of standardized data. The approach can also bring in efficiencies due to automation of mapping and transformation. Some of the features are:

- Automated study mapping using pre-trained TA/indication/project mapping models

- Review of study mapping and continuous model learning based on user feedback
- Programming environment for complex programming scenarios
- Metadata change detection and model recommendation of possible mappings
- Transformation execution using model map's output based on job schedule and on-demand execution
- Automated review of the transformed data using verification tools (e.g. P21 checks)
- Access of the data through web services and data browser through fine grained access data security model

Framework to Build a Clinical Data Virtual Assistant

A virtual assistant enables conversational experience through the following components:

a. Intents – Defines the scope of virtual assistant. The virtual assistant is trained to classify natural language text into pre-defined intents (e.g. adverse event (AE) analysis is an intent for clinical data virtual assistant).
b. Entities – Identifies the information and analysis scope of the intent (e.g. adverse event analysis can be done based on study, site, subject id, body system class, severity – in this case all the above are entities for the adverse event intent).
c. Custom Action – This is a configurable hook for calling end points based on extracted intents and entities.
d. Standardized Data – The custom action integrates with standardized data – in this case, the standardized data is in SDTM format. The data in SDTM format gives significant advantages - the virtual assistant can be plugged into any study (if they are in SDTM format without any set up process).
e. Data Security Access – Rules to be enabled for user, based on role and security group. These rules enable fine grained access for accessing the data to the lowest granular level (such as column/row and value level access).

How is the Conversational Experience Enabled?

Conversational experience is enabled through the following steps executed by the virtual assistant:

a. When the user types in a text, the trained virtual assistant first classifies the text into intent. The system also shows the confidence of the predicted intent and, in case of low confidence, the user can provide feedback on misclassified intent. The feedback will be used for training to improve the accuracy of intent classification.
b. The virtual assistant will also extract the entities involved from the text. The entities in most of the cases are pre-trained (e.g. study names, site names), in some cases they can be trained to be treated as a natural language entity.
c. The virtual assistant invokes the custom action hook based on the intent, passes the entities as parameters, and displays the results as a table, graph and sentence based on the extracted entities.

Following is a table which defines an example of a conversation and the related intents and entities

S.No	Conversational Text	Intents	Entities	Output
1	Hello, can you please give me cardiac disorders related AE for study SA-03-2019-001?	Adverse Event Analysis	SA-03-2019-001 (study) Cardiac Disorders (adverse event body system class)	System displays the cardiac disorders related adverse event in table format
2	Thanks. But can you please filter out only SAEs?	Adverse Event Analysis	Entities from previous conversation: study, adverse event body system class SAE (Serious AE)	System displays the serious adverse events related to cardiac disorders
3	Can you please give me Medical history for the patient SA-03-2019-001-10032-60008?	Medical History	New entity: SA-03-2019-001-10032-60008 (patient)	System displays the medical history details of the patient
4	How about concomitant medications for this patient?	Concomitant Medication	Entities from previous conversation: Patient	System displays the concomitant medication details of the patient
5	Please provide details about drug treatment?	Drug Exposure	Entities from previous conversation: Patient	System provides drug exposure details of the patient
6	I want to check on the patient's HDL, LDL, CHOL lab test values across all visits?	Lab Test Details	Entities from previous conversation: Patient New entity: HDL, LDL, CHOL (lab tests)	System shows the lab values plotted across different visits for HDL, CHOL, LDL lab tests

Table 1. Example of a Conversation and the Related Intents and Entities

How to Extend the Virtual Assistant Capability?

While the virtual assistant is packaged with pre-defined intents, a framework is necessary to extend the virtual assistant scope with additional intents. Steps to extend the virtual assistant are as follows:

a. Intent definition and training intent for intent classification using deep learning models (pre-trained with natural language and clinical vocabulary)
b. Entities identification, extraction from data and entity training
c. Verification of trained model for intent and entities
d. Programming and integration of end points based on entities and intents
e. Deployment of model with entities auto-completion and entities auto-update

Case Study of a Clinical Data Virtual Assistant Proof of Concept:

For the virtual assistant, the following table summarizes the intents and the SDTM domains used in the proof of concept.

S.No	Intent	Intent Definition	SDTM domains
1	Studies in System	Gives detail of configured studies	NA
2	Site and Investigator Name	Gives detail of Sites and investigator names in the Study	DM (Demographics)
3	Patient Demographics Detail	Shows patient details by entities such as age, race, gender, site and country	DM (Demographics)
4	Patient Adverse Event Detail	Shows adverse event details by entities such as age, race, gender, site, body system, SAE and country	AE (Adverse Events) , DM (Demographics)
5	Patient Medical History	Shows medical history of patient	MH (Medical History)
6	Patient Exposure Details	Shows drug exposure details for a patient	EX (Exposure)
7	Patient Concomitant Medication Details	Shows patient's concomitant medication detail	CM (Concomitant Medications)
8	Patient Lab Test Details	Shows patient's lab test details	LB (Laboratory test Results)
9	Patient Vital Sign Details	Shows patient's vital sign details	VS
10	Patient's ECG Details	Shows patient's ECG details	EG

Table 2. Virtual Assistant Intents and SDTM Domains in Proof of Concept

The proof of concept involved ingesting raw EDC data and transforming them into standardized data through a metadata transformation approach (which can be invoked on a demand basis). The standardized data was integrated with clinical data virtual assistant to enable conversational experience (CX) on clinical data.

CONCLUSION

A virtual assistant trained on natural language and clinical standards (such as SDTM) offers a new, efficient way of interacting with clinical trial data. It enables end users such as data managers, biostatisticians and medical monitors to get clinical trial insights through conversations, instead of navigating through multiple reports and dashboards, and manually deriving the insights from the collated information. This solution can be extended for all the patient data domains and can be a unified solution to cater to clinical trial analytics needs across multiple clinical trial stakeholders. While the virtual assistant can be integrated with direct upload of SDTM data, the true value of the clinical trial virtual assistant is realized when integrated with a modern clinical data analytics platform. Such a next-generation clinical data analytics platform, enabled through metadata and machine learning approaches, provides near real-time standardized data. Standardized data generated on demand enables a virtual assistant to provide conversational experience for providing rich, meaningful, faster and more effective clinical trial insights.