Data Quality Evaluation of EHR FHIR API Real World Data for FDA & Pharmaceutical Industry Use for Analysis and AI Implementation

Sarah Ferko, IBM, Washington, DC, USA
Abhishek Pandey, IBM, Washington, DC, USA
Matthew Deady, IBM, Washington, DC, USA
Brian Goodness, IBM, Washington, DC, USA

ABSTRACT

The FDA and pharmaceutical industry's pre-market and/or post-market activities could possibly be improved by using real-world data (RWD) from healthcare organizations' electronic health record (EHR) databases and application of artificial intelligence (AI) tools. In the U.S., this aim has been traditionally hindered by the inability to exchange clinical data between healthcare providers, pharmaceutical companies, regulatory agencies, and public health organizations in an interoperable format. However, the recently enacted 21st Century Cures Final Rule requires all healthcare providers to support seamless and secure access, exchange, and use of EHR data through an interoperable FHIR® (Fast Healthcare Interoperability Resources) standard API endpoint. This paper describes methods and tools to analyze the data quality and interoperability from these API endpoints to evaluate them as possible sources for clinical trial data collection and where AI algorithms may be used to enhance this process. The paper identifies several CDISC data requirements that can be mapped to the OHDSI Data Quality Dashboard data quality tests and identifies high-level areas for potential concern.

INTRODUCTION

Pharmaceutical companies utilize EHR data to enhance their data sources within a clinical trial and provide a real-world perspective on the safety and efficacy of a medical product. EHR data can provide a comprehensive and real-world view of health history, treatment outcomes, and medical conditions on a diverse patient population, resulting in a larger, more representative sample of patients. Within a clinical trial, EHR data may be leveraged as a historical control arm or used to conduct comparative effectiveness research, where a pharmaceutical company uses EHR data to compare their drug's performance against other available treatments. EHR data may also be used for long-term safety monitoring, which is conducted post-approval and allows pharmaceutical companies to identify and assess safety profiles over a longer term to ensure ongoing safety of the product. Additionally, pharmaceutical companies may use EHR data to support label expansions and post-marketing commitments. To increase the efficiency of this RWD, AI tools can be applied on EHR data to more effectively tabulate unstructured data from clinical notes, resulting in more comprehensive and robust data. Ultimately, the use of EHR data paired with AI enhancements can contribute to accelerating the drug development process, improve trial outcomes, and overall contribute to advancing healthcare research.

As part of its work for the FDA's Biologics Effectiveness and Safety (BEST) initiative, IBM built a pilot platform that allows the FDA to receive RWD from EHRs to improve their post-market surveillance capabilities. This platform provides the FDA with the critical data elements not easily found in claims data or information that may only be found in clinical notes for patients with probable adverse events, such as vaccine administration and clinical note narratives. A similar platform or method could be used to assist pharmaceutical companies, provided they have the legal permission to view the patient data, in receiving RWD when used in a clinical trial submission or post-market commitments. The platform takes advantage of the Office of the National Coordinator (ONC) Cures Act Final Rule, which supports seamless and secure access, exchange, and use of Electronic Health Information (EHI) by requiring certified EHR to support Fast Healthcare Interoperability Resources (FHIR) as the federally mandated standard (1).

The HL7® FHIR® standard defines a set of basic building blocks, called resources, which are a generic definition of common health care concepts (e.g., patient, observation, practitioner, device, condition) (2). FHIR was developed by Health Level Seven, Inc. (HL7) (2) to facilitate interoperability between health systems (3) and was designed to improve existing standards by reducing implementation complexity without losing information integrity (2) (4). Seamless access is achieved by requiring data to be interoperable, which is defined as two or more software or systems with the ability to read and make use of the information received from one another. To better define the standard and its implementation, the ONC for Health Information Technology released a set of minimum requirements known as the United States Core Data for Interoperability (USCDI) (5). The Argonaut Project FHIR accelerator group has published the U.S. Core implementation guide (IG) that describes how to exchange USCDI data using FHIR (6). As of December 31, 2022, the ONC enacted the requirement that EHR vendors must provide FHIR application programming interfaces (APIs) supporting the first versions of the USCDI (v1) and U.S. Core (v3.1.1) to all customers or be at risk of losing their Healthcare IT developer certification (7).

Given this newness of the ONC's regulation, there are open questions about the feasibility of using this EHR data for

clinical trial submissions. These include determining the quantity and type of data available and whether the data elements to be used in clinical trial submissions to the FDA are sufficiently populated and/or included in current or future versions of the USCDI / U.S. Core requirements. To answer these questions, a study can be conducted, which is described in detail in the sections below. Any gaps identified by the data quality analysis can help inform the FDA and the healthcare informaticist community about the current state of FHIR data exchange for a public health use case, given the way in which current and future planned versions of the USCDI and/or U.S. Core requirements are being implemented. These analyses are also expected to provide recommendations for changes in future versions of the USCDI / U.S. Core requirements so that the RWD can be made readily available to public health authorities.

Once the data quality is confirmed, additional AI methods may be needed to summarize the data collected given the volume per patient case. IBM developed a generative AI patient summarizer tool, an application of the retrieval-augmented generation (RAG) AI framework (8). This application uses a foundation model to extract medical entities of interest from a corpus of unstructured clinical notes. The patient summarizer can be used in clinical trials to summarize each patient's case data, including both structured and unstructured data. For unstructured data, the tool automatically extracts the information of interest from the text. Within a clinical trial, this tool can be expanded to tabulate the unstructured data by data element. Many important data elements needed for analysis may be submitted within an unstructured clinical note rather than as structured data. For example, this may include death information (e.g., cause(s) of death, primary) or condition information (e.g., whether an action was taken with the treatment such as a dose modification or pause in treatment, whether the treatment resulted in death or hospitalization).

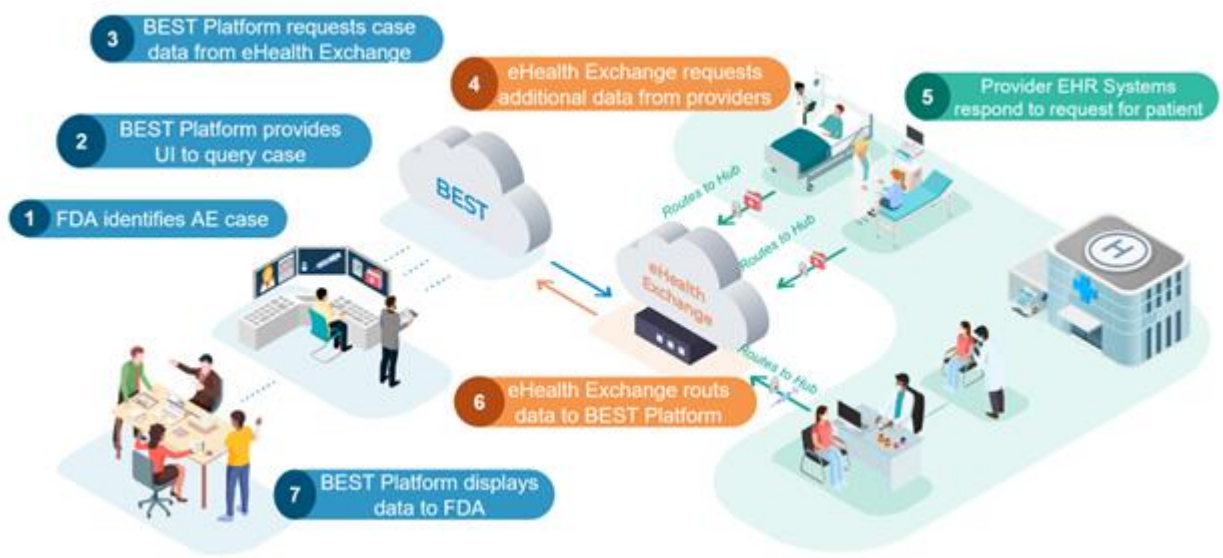
METHODS

This section describes the methods for a data quality study to assess the overall fitness of RWD from EHR's FHIR APIs for use in clinical trial submission received through a platform like the FDA BEST platform.

DATA

The RWD to be assessed can be obtained through a platform like FDA BEST described above. FDA BEST (Figure1) collects data for patient vaccine adverse event cases from participating healthcare organizations through the eHealth Exchange (eHX). eHX is a Health Information Exchange (HIE) network that covers 77% of all U.S. hospitals, 61 state and regional health information exchanges, and five federal agencies. It uses a common set of standards and specifications to establish a trusted, interoperable connection to securely share detailed clinical information (9).

FIGURE 1: FDA BEST PILOT PLATFORM



DATA QUALITY ASSESSMENT

To assess whether the RWD data delivered to the platform was of sufficient quality for clinical trial submission, we would suggest using the framework for evaluating clinical data quality suggested by Kahn, et al. (10), which is well established, used throughout the industry, and recommended by clinical research groups such as Observational Health Data Sciences and Informatics (OHDSI) (11) and Patient-Centered Outcomes Research Institute (PCORI) (12). Table 1 gives an outline of the framework.

TABLE 1: DATA QUALITY FRAMEWORK

Conformance	Completeness	Plausibility
Do data values adhere to specified standards and formats? Sub-types include Value, Relational, and Computational. *Relevant Example: Are resources coded using interoperable code systems?*	Are variables present and do they contain all recorded values? *Relevant Example: Is prescribed dosage and route captured for all medications?*	Are data values believable? Sub-types include Uniqueness, Atemporal, and Temporal. *Relevant Example: Does the AE date occur after the patient's birth date?*

The Kahn et al. framework does not prescribe individual data tests but describes the types of questions that fit under each topic of conformance, completeness, and plausibility. To develop our unique data quality tests, we created an initial pool based on those implemented by the open source OHDSI Data Quality Dashboard (DQD) tool (13) or automated data quality assessment. The DQD tool has a suite of tests based on the Kahn et al. framework used to assess a data set stored in an Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We translated these DQD tests, when applicable, to assess the same data quality question in FHIR format. From this initial group of tests, we selected a subset that matched similar elements or data conformance rules needed for clinical trial submission as defined by the CDISC SDTM and SDTMIG conformance rules (14). This subset focused on rules around medication exposure, condition diagnosis, lab test results, and patient demographics. Data elements tested were selected without consideration for if they are required by current USCDI / U.S. Core requirements. Thus, when data elements or conformance specifications are tested that are not included in the current USCDI / U.S. Core requirements, it would not be surprising to see a failing test result, which does not reflect upon the USCDI/U.S. Core compliance of EHR vendors or participating data partners.

FHIR search requests were developed to query the platform's FHIR server to obtain results for all data quality tests. The FHIR search is the primary mechanism used to find and list resource instances in the FHIR specification (15). Test results would report the percentage of resources fulfilling the test criteria out of all applicable resources. These tests describe the existing complexity within RWD exchanged in FHIR format for a clinical trial use case based on collection of data quality checks applied to a subset of domains submitted within an NDA/BLA submission. Beyond tests for conformance and completeness, additional tests for plausibility are examined when applicable, given that EHR data can be prone to errors not often observed within RCT data given the controlled nature of the study (e.g., treatment dates occurring before a patient's birth date or after a patient's death date). The FHIR search queries for each test can be shared upon request. Test results <100% do not indicate a failure to meet a standard, since we did not define an a priori level of completeness to pass the test given the many valid reasons or specific allowances for a missing data element, expected format, etc., in FHIR data.

GENERATIVE AI PATIENT SUMMARIZER TOOL

Previous research documents the amount of RWD is exploding through the increasing use of EHRs (16). High data volume per patient case will lead to challenges in clinical analysis of received EHR data. These challenges will be especially pronounced when analyzing unstructured text of clinical notes which are estimated to be ~80% of all clinical data (17). Extracting information from unstructured text is an excellent use case for a generative AI tool that can collect the pertinent evidence and summarize all the patient's case data (structured and unstructured). To investigate the value, we built a generative AI patient summarization tool using the Retrieval Augmented Generation (RAG) AI design pattern (8). RAG applications take advantage of the recent advances in Large Language Models (LLMs), like ChatGPT, but are built to answer questions on the content in a corpus of documents that may not be in the LLM's training set. This could include very recent information, expert information on obscure subjects or private or proprietary information such as a patient's health record. RAG uses a semantic search to identify relevant documents in the corpus and then feeds that context along with the original question to the LLM. Our application ran the LLM locally as we downloaded to our cloud environment several widely used open source LLMs such as Meta Llama2 7b, Google's Flan T5 and ChatGPT-3.5. To prevent common issues with LLM including hallucinations and lack of explainability, we tested whether our RAG application could link all generated content to the source text, identified through the aforementioned semantic search, for easy review of correctness and context.

Our generative AI patient summarizer app searches through the corpus of text which are the unstructured clinical notes found in DocumentReference resources from the health provider FHIR APIs. The tool first summarizes the text from tens to hundreds of clinical notes per patient into a single paragraph of key content. Next, it creates text embeddings for each note so that it can run a semantic search for medical entities potentially relevant to creating an adverse event report for a clinical trial submission. These entities in our initial version are allergies, medical conditions, and medications, but can be updated as necessary. Each entity extracted links back to the source clinical note (and displays the date) from where the evidence was found allowing the preparer of the clinical trial data submission package to read any additional information or context around the extracted information and to confirm that it is not an LLM-generated hallucination. This demo was created using the clinical notes from Synthea synthetic data from MITRE (18) which are captured in FHIR DocumentReference resources. This is the same method that clinical note data is exchanged between healthcare entities on a HIE. If real patient notes are used, the proper HIPAA protocols must be observed and the LLM should be downloaded to a HIPAA compliant environment.

RESULTS

Demographics for the patients in the EHR should be collected to assess the contents of the data being analyzed for quality. Table 2, a summary of 300 synthetic patients, shows what could be found in EHR data, grouped by predefined age groups and nationally defined systems for gender, race, and ethnicity.

TABLE 2: DEMOGRAPHICS OF SYNTHETIC PATIENT COHORT

Category	Demographic group	Total Patients	
		N	%
Total	**Total**	**300**	**100.0%**
Age	Under 5 years	18	6.0%
	5 to 17 years	17	5.5%
	18 to 24 years	17	5.5%
	25 to 44 years	67	22.6%
	45 to 64 years	82	27.4%
	65 years and above	99	33.0%
	Missing	0	0.0%
Gender	Male	75	25.0%
	Female	225	75.0%
	Unknown	0	0.0%
Race	White	160	53.5%
	Black or African American	46	15.5%
	Asian/Pacific Islander	17	5.5%
	American Indian or Alaska Native	5	1.5%
	Other	33	11.0%
	Unknown	36	12.0%
	Declined to Answer	3	1.0%
Ethnicity	Hispanic	45	15.0%
	Non-Hispanic	227	75.8%
	Unknown	28	9.2%

DATA QUALITY ANALYSIS

We identified 30 data quality tests to assess completeness, conformance, and plausibility of the FHIR data for CDISC data requirements. The tests are listed for demographics, medication exposure, adverse event information, and lab test results in Tables 3-6 below:

TABLE 3: DEMOGRAPHIC DATA QUALITY TESTS

Category	Test Name
Completeness	Completeness – Patient birth date
Completeness	Completeness – Patient gender
Completeness	Completeness – Patient race
Completeness	Completeness – Patient race collected
Completeness	Completeness – Patient ethnicity
Completeness	Completeness – Patient ethnicity collected
Conformance	Conformance – Patient expected gender value
Conformance	Conformance – Patient expected race value
Conformance	Conformance – Patient expected race value set
Conformance	Conformance – Patient expected ethnicity value

Category	Test Name
Conformance	Conformance – Patient expected ethnicity value set
Plausibility	Plausibility – Patient birth date > 1850
Plausibility	Plausibility – Patient birth date should b before death date
Plausibility	Plausibility – Patient birth date should be before today's date

TABLE 4: MEDICATION EXPOSURE DATA QUALITY TESTS

Category	Test Name
Completeness	Completeness – MedicationRequest medication
Completeness	Completeness – MedicationRequest prescribed date
Completeness	Completeness – MedicationRequest status
Completeness	Completeness – MedicationRequest reason
Completeness	Completeness – MedicationRequest code
Conformance	Conformance – MedicationRequest Medication non-local code system
Plausibility	Plausibility – MedicationRequest prescribed date > 1950
Plausibility	Plausibility – MedicationRequest prescribed date before birth date
Plausibility	Plausibility – MedicationRequest prescribed date after death date
Plausibility	Plausibility – MedicationRequest prescribed date after today's date

TABLE 5: ADVERSE EVENT DATA QUALITY TESTS

Category	Test Name
Completeness	Completeness – Condition encounter diagnosis code
Completeness	Completeness – Condition encounter diagnosis abatement date
Completeness	Completeness – Condition encounter diagnosis body
Completeness	Completeness – DocumentReference clinical note data successfully retrieved
Conformance	Conformance – Condition encounter diagnosis valid encounter reference
Conformance	Conformance – Condition non-locally defined code system
Plausibility	Plausibility – Condition encounter diagnosis onset date > 1950
Plausibility	Plausibility – Condition encounter diagnosis abatement date > 1950
Plausibility	Plausibility – Condition encounter diagnosis onset date < birth date
Plausibility	Plausibility – Condition encounter diagnosis abatement date < birth date
Plausibility	Plausibility – Condition encounter diagnosis onset date > death date
Plausibility	Plausibility – Condition encounter diagnosis abatement date > death date
Plausibility	Plausibility – Condition encounter diagnosis onset date > today's date
Plausibility	Plausibility – Condition encounter diagnosis abatement date > today's date

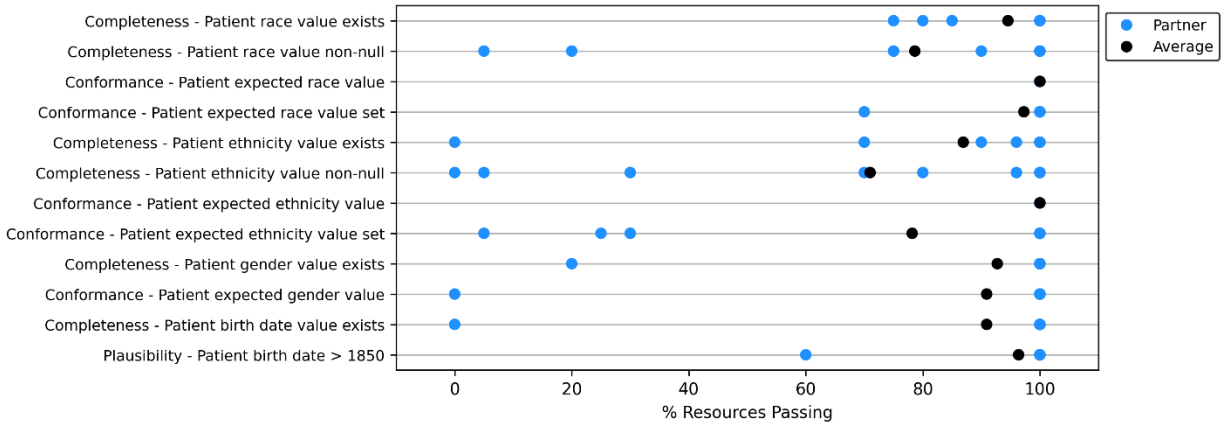
TABLE 6: LAB TEST RESULTS DATA QUALITY TESTS

Category	Test Name
Completeness	Completeness – Lab Test Observation code
Completeness	Completeness – Lab Test Observation status
Completeness	Completeness – Lab Test Observation value
Completeness	Completeness – Lab Test Observation specimen
Conformance	Conformance – Lab Test Observation expected status value
Conformance	Conformance – Lab Test Observation code system
Plausibility	Plausibility – Lab Test Observation date > 1950
Plausibility	Plausibility – Lab Test Observation date < patient's birth date

Category	Test Name
Plausibility	Plausibility – Lab Test Observation date > patient's death date
Plausibility	Plausibility – Lab Test Observation date > today's date

Example results for synthetic data for a subset of tests relevant to CDISC patient demographics data are displayed in Figure 2. The figure demonstrates the range of theoretical results for percent of resources that passed the test for each partner (light blue) and the average for all partners (dark blue).

FIGURE 2: SYNTHETIC DATA QUALITY RESULTS FOR PATIENT DEMOGRAPHIC RESOURCES



GENERATIVE AI PATIENT SUMMARIZER TOOL

Given that the patient summarizer tool is still in its prototype phase, it has not been deployed to any environment with real patient data. Thus, no study was able to be conducted to assess the effectiveness of this tool. However, the tool was used to review the functionality with our team's clinicians and received positive feedback. Representative screen shots of the software are included in Figures 3 and 4 below.

FIGURE 3: SCREENSHOT OF RAW CLINICAL NOTE TEXT DATA FOR SYNTHETIC PATIENT LOADED INTO PATIENT SUMMARIZER UI

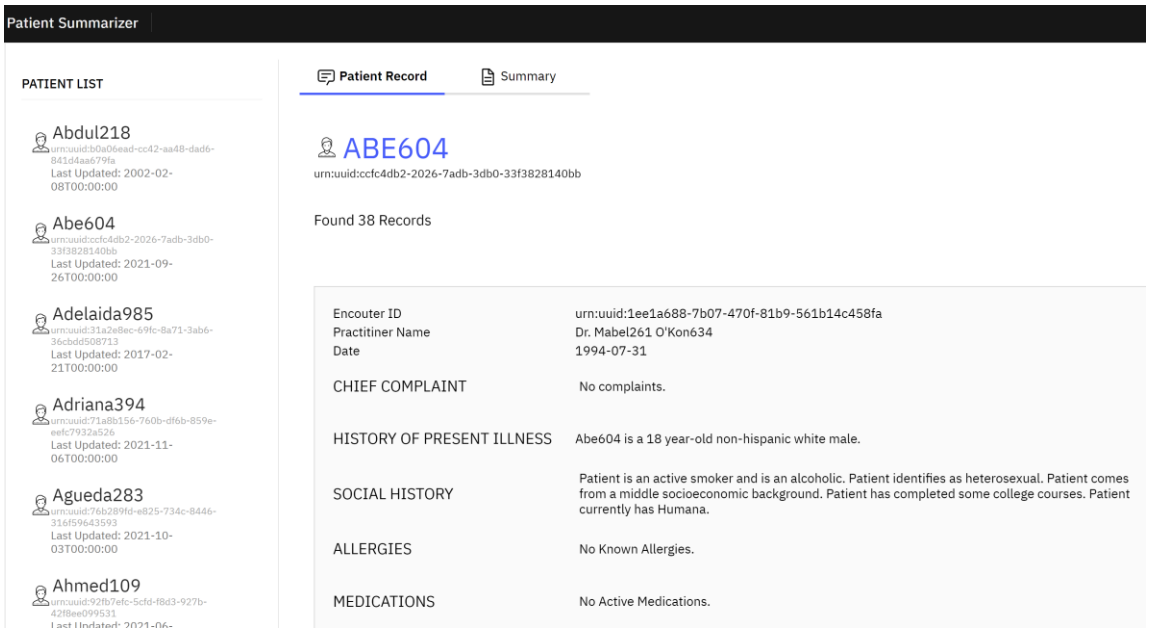
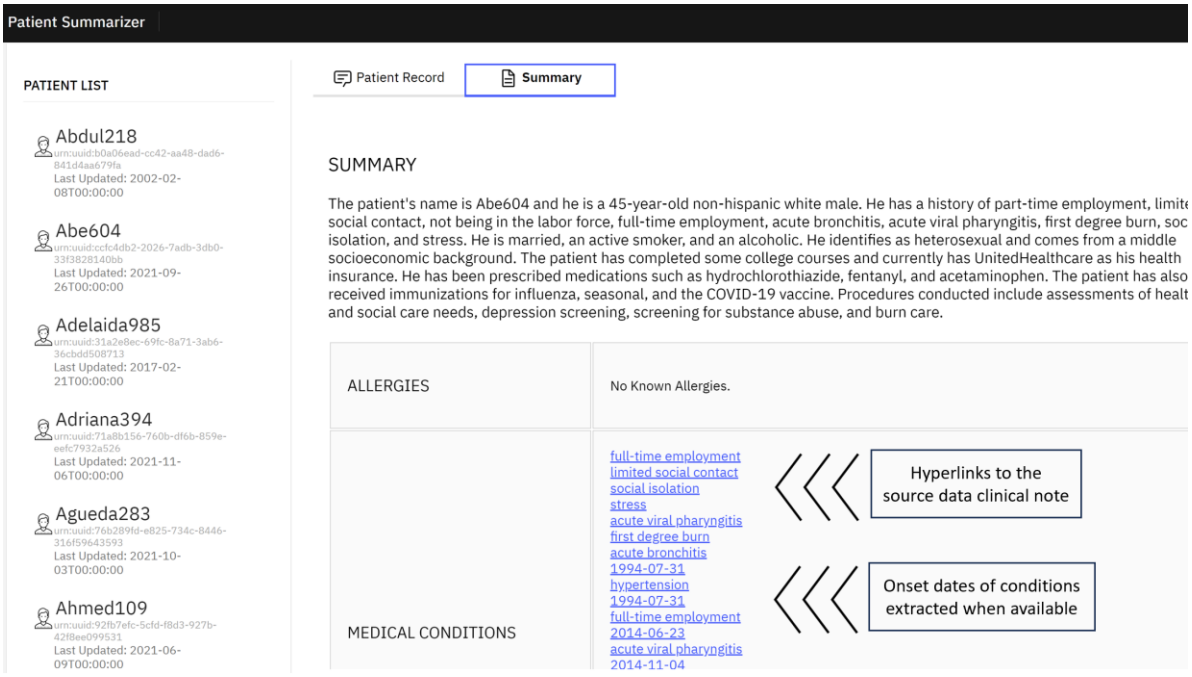


FIGURE 4: SCREENSHOT OF PATIENT SUMMARIZATION TOOL RESULT FROM THE SUMMARIZATION OF SYNTHETIC PATIENT'S 38 CLINICAL NOTES



CONCLUSION

This study describes several useful methods and tools relevant for any organization that intends to receive and use patient healthcare data from a FHIR API, including for use by pharmaceutical companies to support FDA clinical trial submissions. Given the generally accepted messiness of EHR data, having methods to evaluate the data quality and extract important information only found in unstructured clinical notes is extremely important for this task. Several high-level data known issues with EHR FHIR data are discussed in more detail below:

DATA ELEMENT NOT REQUIRED BY USCDI

Many of the CDISC data elements used for clinical trials are not listed as "Mandatory" or "Must Support" by U.S. Core and are not supported by Epic FHIR APIs. It is promising to see that Epic's FHIR API documentation shows that it supports some elements not required by USCDI, and U.S. Core, such as prescribed dosage timing, medication administration route and lab test specimen reference. However, many CDISC required fields are not, including reason medication given, and reason concomitant/prior medications were discontinued or had their dosage adjusted. An alternative method to obtain an approximation of the information needed for these data elements that are not supported is through the clinical notes captured in DocumentReference resources which include narratives about the case that can fill in many of the blanks left by various missing data elements.

USE OF LOCALLY DEFINED, NON-INTEROPERABLE CODE SYSTEMS

One of the key challenges to healthcare data exchange is achieving semantic interoperability, when data are not only exchanged between two systems but also understood by each system. While the U.S. Core requirements prescribe standard code or value sets for most of the applicable data elements in their profiles, they prioritize requiring standardized code sets for the coded value of a resource. This functionality allows a clinician to understand specifics of diagnosis, medication, lab test, or other events of the resource.

The requirements have larger conformance problems for the non-code data elements, including some in which the current U.S. Core specification only suggests, but does not require, a standard code or value set to use. This is due to a systemic issue across the industry which has not defined, specified, or coalesced around standard value sets or code lists for these elements. These specifications allow for some or all of the codes to be locally defined values/codes through value sets that are extensible (17, preferred (21) or example (22). Example sets are especially concerning given the implementer is not expected or even encouraged to draw from the specified values (22). For elements such as encounter care setting, immunization site, and route, because there is no requirement to match a specified code or value set, they are usually populated with codes defined locally at the EHR site. Each of these fields has a text description that can be easily understood by a clinician, but the lack of a standardized code system can cause difficulties for any simple machine processing or algorithmic use of the data. Future research could analyze the ability of natural language processing (NLP) techniques to identify similar values across data partners and map them to an expected code system or value set. Additionally, future versions of the USCDI or U.S. Core could change the requirement to defined required code lists, using published, standardized code lists or value sets.

MISSING MEDICATION COMPLIANCE DATA

Because the Medication domain in the FHIR standard includes a number of related resources it is difficult to assess whether the necessary data elements are populated. Some medication resources that might apply to clinical trial data include:

- **MedicationRequest:** Used for an order for both supply of the medication and the instructions for administration of the medicine to a patient (23).
- **MedicationDispense:** The provision of a supply of a medication with the intention that it is subsequently consumed by a patient (usually in response to a prescription, order, or request) (24).
- **MedicationAdministration:** When a patient consumes a medicine, or it is otherwise administered to them.
- **MedicationStatement:** A record of medication being taken by a patient, or that the medication has been given to a patient where the record is the result of a report from the patient, or another clinician.

Currently, the required version (3.1.1) of the U.S. Core mandates that MedicationRequest resource be populated but amends the definition of MedicationRequest to be able to return all prescribed and "self-prescribed" medications directly ordered by a provider or reported by the provider, patient, or related person including ethical drugs, over the counter (OTC) medication, and other substances taken for medical and/or recreational use. This ensures "self-prescribed" medications are captured in MedicationRequest as opposed to other resources, such as MedicationStatement. However, this definition does not convey compliance to the prescription (24), so in cases that the patient does not comply (or does not comply at the prescribed time), it can be difficult to know exactly what other medications were being taken at the time of administration. Future versions of the U.S. Core include the MedicationDispense profile (25), which could address concerns about including medication compliance information and better support this FDA use case as well as other drug safety and effectiveness use cases.

ENHANCEMENTS FOR FUTURE USE IN CLINICAL TRIALS

The data quality test results highlighted within this paper focus on a subset of checks within a subset of domains: Demographics, Exposure, Laboratory Results, and Adverse Events to illustrate their overlap with existing checks and utility within the clinical trial setting when EHR data is used. In addition to data quality checks related to conformance to CDISC and completeness, which may already be regularly checked by existing Commercial Off-the-Shelf Software (COTS) or by internal pharmaceutical company tools, this paper introduces the need for additional data quality checks related to plausibility, which is strongly encouraged when using EHR data in clinical trials due to the plausibility errors that can be present within the data. Thus, the data quality checks presented in this paper identify those that overlap with existing checks around conformance and completeness and those that are needed in addition to existing checks. However, there are existing conformance and completeness checks that are not yet coded within the EHR data quality checks but are needed to support a regulatory submission in CDISC format. To fully leverage these checks within a regulatory submission, these checks would need to be added. Checks would also need to be expanded to all SDTM domains submitted within study in support of an NDA/BLA, beyond the four domains listed above. As indicated within the FDA Data Standards Catalog (19), submitting data using CDISC data standards is required at the time of this paper's publication. Thus, these checks would need to be expanded to support the SDTM and SDTMIG Conformance Rules published by CDISC (20) in addition to completeness checks based on the SDTM Core Designations published within the SDTMIG (e.g., Required, Expected, Permissible).

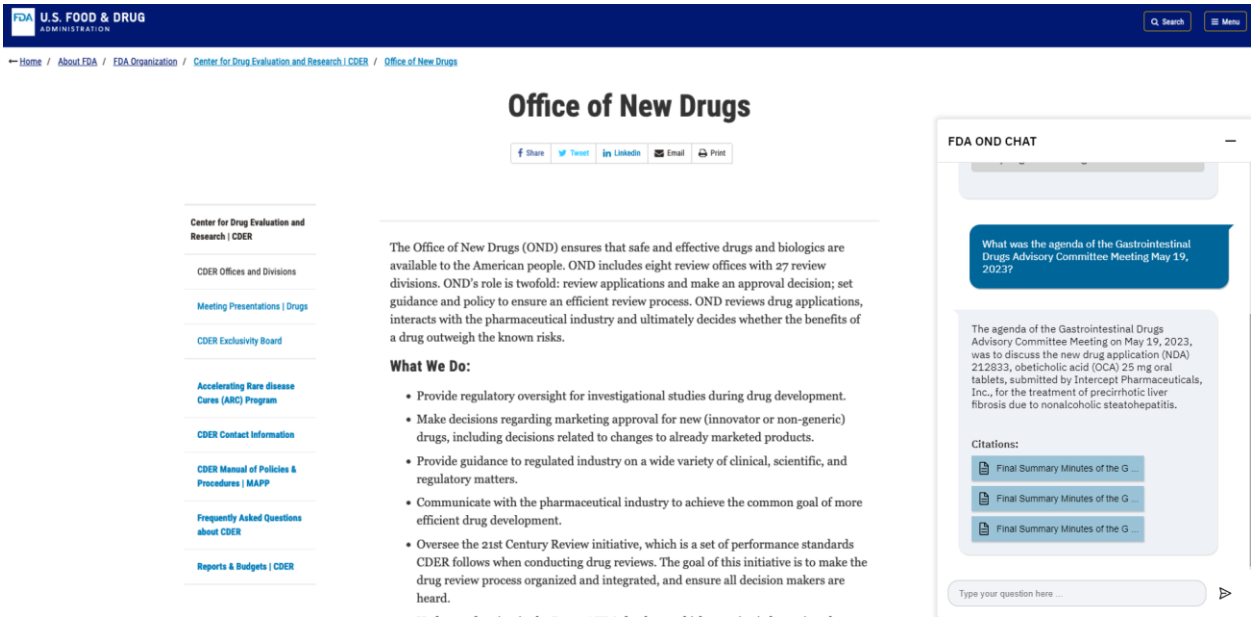
FURTHER DEVELOPMENT AND USE OF DATA QUALITY CHECKS IN CLINICAL TRIALS

As presented in the Methods section, to fully leverage the EHR data quality checks to support a RWD study as part of an NDA/BLA, checks will be expanded beyond the initial domain areas explored for this study: Demographics, Exposure, Laboratory Results, and Adverse Events, and additional relevant checks for plausibility will also be added. In addition, existing completeness and conformance checks currently used in regulatory submissions to assess CDISC data will be included.

GENERATIVE AI PATIENT SUMMARIZER TOOL: RESULTS AND FURTHER DEVELOPMENT

Feedback about the tool suggests that Generative AI & NLP patient summarization is a promising avenue for future research. The tool can be helpful to uncover and confirm necessary information that is only contained in unstructured clinical notes, which may be necessary for a pharmaceutical company to describe adverse events that occurred to a patient. IBM plans to continue to investigate the ability of RAG to be able to be able to assist in clinical review and summarization of a patient clinical notes by adding additional functionality to the tool, testing out additional open source LLM models and testing on real patient clinical note data. We hope to publish our findings to encourage the community, if we show successful result, to develop their own RAG implementations using open source LLMs. Future research for this method can also investigate the ability to answer specific questions (with links to source data) about content of all the RWD EHR data for a patient. See Figure 5 below for an example screenshot of a RAG application that answers questions on results of FDA Office of New Drug advisory committee meetings and other FDA guidance.

FIGURE 5: SCREENSHOT OF CHATBOT TO ANSWER QUESTIONS ON FDA OND ADVISORY COMMITTEE MEETING GUIDANCE AND OTHER GUIDELINES WITH CITATION LINKS



REFERENCES

1. ONC. Cures Act Final Rule. 2023. Available from: https://www.healthit.gov/topic/oncs-cures-act-final-rule
2. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. JMIR Med Inform. 2021 Jul 30;9(7):e21929. eng. Conflicts of Interest: None declared. Epub 20210730. doi:10.2196/21929. Cited in: Pubmed; PMID 34328424.
3. Health Level Seven International. Welcome to FHIR (Fast Healthcare Interoperability Resources). 2023. Available from: https://fhir.org/
4. Health Level Seven International. Introducing HL7 FHIR (Fast Healthcare Interoperability Resources) 2023. Available from: https://www.hl7.org/fhir/summary.html.
5. Office of the National Coordinator for Health Information Technology. Cures Act Final Rule: Standards-based Application Programming Interface (API) Certification Criterion. 2022 100423. Available from: https://www.healthit.gov/cures/sites/default/files/cures/2020-03/APICertificationCriterion.pdf.
6. Health Level Seven International. US Core Implementation Guide. 2023. Available from: https://build.fhir.org/ig/HL7/US-Core/index.html.
7. Shanbhag A, Anthony R. On the Road to Cures Update: Certified API Technology. healthIT.gov; 2022. Available from: https://www.healthit.gov/buzz-blog/healthit-certification/on-the-road-to-cures-update-certified-api-technology.
8. What is retrieval-augmented generation? IBM Research Blog. 2021. Available from: https://research.ibm.com/blog/retrieval-augmented-generation-RAG
9. Exchange e. Health Information Exchange. 2022. Available from: https://ehealthexchange.org/.
10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244. eng. Epub 20160911. doi:10.13063/2327-9214.1244. Cited in: Pubmed; PMID 27713905.
11. Schuemie M, Huser V, Blacketer C. Chapter 15 Data Quality. Book of OHDSI (Observational Health Data Sciences and Informatics). 2021.
12. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, Wu Y, Prosperi M, George TJ, Harle CA, Shenkman EA, Hogan W. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Dec 9;27(12):1999-2010. eng. doi:10.1093/jamia/ocaa245. Cited in: Pubmed; PMID 33166397.
13. Observational Health Data Sciences and Informatics. Data Quality Dashboard. 2023. Available from: https://github.com/OHDSI/DataQualityDashboard.
14. CDISC. SDTM and SDTMIG Conformance Rules v2.0. 2021. Available from: https://www.cdisc.org/standards/foundational/sdtmig/sdtm-and-sdtmig-conformance-rules-v2-0.
15. Health Level Seven International. FHIR CI-Build Search. 2023. Available from: https://build.fhir.org/search.html#:~:text=FHIR%20Search%20is%20the%20primary,enough%20to%20be%20commonly%20useful.
16. Kong HJ. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res. 2019 Jan;25(1):1-2. doi: 10.4258/hir.2019.25.1.1. Epub 2019 Jan 31. PMID: 30788175; PMCID: PMC6372467.
17. Health Level Seven International. Release 4 Terminologies -Extensible. 2019. Available from: https://hl7.org/fhir/R4/terminologies.html#extensible.
18. 27. MITRE. SyntheticMass Downloads. Available from: https://synthea.mitre.org/downloads

19. FDA. 2023. Data Standards Catalog. 2023. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog

20. CDISC. SDTM. 2021. Available from: https://www.cdisc.org/standards/foundational/sdtm

21. Health Level Seven International. Release 4 Terminologies - Preferred. 2019. Available from: https://hl7.org/fhir/R4/terminologies.html#preferred.
22. Health Level Seven International. Release 4 Terminologies - Example. 2019. Available from: https://hl7.org/fhir/R4/terminologies.html#example.
23. Health Level Seven International. Fast Healthcare Interoperability Resources, US Core Implementation Guide, Medication Request. 2023. Available from: https://build.fhir.org/medicationrequest.html.
24. Health Level Seven International. Fast Healthcare Interoperability Resources, US Core Implementation Guide, Medication Dispense. 2023. Available from: https://build.fhir.org/medicationdispense.html
25. Health Level Seven International. Options for Representing Medication. 2020. HL7 FHIR® US Core Implementation Guide STU3 Release 3.1.1. Available from: https://hl7.org/fhir/us/core/STU3.1.1/all-meds.html#options-for-representing-medication.

ACKNOWLEDGMENTS

Development of the initial study this paper is based on benefitted from significant engagement with FDA CBER team members and their partners. We thank them for their contributions. The description of eHealth Exchange and the ability to receive the data for the assessment would not have been possible without help from Mike McCune, Kathryn Bingman and Eric Heflin from eHealth Exchange.

CONTACT INFORMATION

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

Author Name: Matthew Deady
Company: IBM
Address: 1512 K St. SE, Unit 4, Washington, DC 20003
Work Phone: 585-820-3425
Email: mjdeady@us.ibm.com