HL7 FHIR Synthetic Data Event Summary Report

August 29 – 30, 2023

HL7 International

Table of Contents

Overview

The HL7® FHIR® Synthetic Data Event was a virtual summit that brought together the academic, standards, and vendor communities for an interactive dialog about creating and using openly accessible synthetic FHIR data for education, standards development, and testing. It targeted those involved in the development, implementation, and/or testing of HL7 FHIR standards and artifacts, including health IT vendors, healthcare payers, standards developers, and government agencies. Over the course of two days, Event participants gained:

- An understanding of the current landscape of synthetic data repositories and tools;
- Insights into the limitations of current synthetic data and tools versus community needs; and
- Methodologies for generating openly accessible synthetic FHIR data that is scalable, extensible, and fit for purpose.

Two sequential tracks focused on sharing information and demonstrating tools and resources that are crucial to the continued expansion and support of the FHIR implementation and testing community. The Event featured individual presentations with hands-on components and panel sessions that encouraged participants to actively engage with the panelists.

The Event convened 28 speakers and moderators, 3 ONC staff, and 69 paid registrants. Attendees represented 59 different organizations from 12 countries.

Speakers and Topics

Opening and Closing Remarks

Topic	Speakers and Affiliations
Day 1: Welcome and Opening Remarks	Daniel Vreeman, HL7
	Kenneth Wilkins, PhD, National Institutes of Health (NIH)
Day 1: Wrap-Up	Daniel Vreeman, HL7
Day 2: Welcome and Announcements	Diego Kaminker, HL7
Day 2: Closing Remarks	Viet Nguyen, HL7

Track 1: The Current Landscape of Synthetic Data Repositories and Tools

Three speakers addressed the goals, scope, and history of existing synthetic data tools and data sets, and their relationship with FHIR.

Topic	Speakers and Affiliations
Scope and History of Synthea	Jason Walonoski, MITRE
Scope and History of MIMIC IV	Alistair Johnson, The Hospital for Sick Children
Synthetic Healthcare Database for Research (SyH-DR)	Zeynal Karaca, Agency for Healthcare Research and Quality (AHRQ)

Track 2: Community Needs and the Limitations of Current Synthetic Data and Tools

Four sessions on Day 1 and four sessions on Day 2 featured moderated panel discussions, each focused on a practical topic addressing barriers and successful use of synthetic data for specific applications.

Topic	Speakers and Affiliations
Barriers and Limitations of Current Tools	Moderator: Lenel James, Blue Cross Blue Shield Association (BCBSA) Panelists: • Jason Walonoski, MITRE • Alistair Johnson, The Hospital for Sick Children • Zeynal Karaca, AHRQ
Synthetic Data for Testing	Moderator: Viet Nguyen, HL7 Panelists: • Adam Wilcox, Washington University School of Medicine • Robert Scanlon, MITRE • Bryn Rhodes, Smile Digital Health • Joseph LeGrand, Vanderbilt University Medical Center/CodeRx
Fairness in Synthetic Data	Moderator: Daniel Vreeman, HL7 Panelists: • Karan Bhanot, Rensselaer Polytechnic Institute • Barbara Draghi, Medicines and Healthcare Products Regulatory Agency • Gino Tesei, Elevance Health, Inc.
A Use of Synthetic Data in Education	Moderator: James Hellewell, Intermountain Health Panelists: • Randi Foraker, Washington University in St. Louis School of Medicine • Mark Braunstein, MD, Georgia Institute of Technology
A Trove of Synthetic Data, But No FHIR Access	Moderator: Diego Kaminker, HL7 Panelists: • Eric Pan, Stanford • Michael Riley, Georgia Institute of Technology
Managing “Big” Synthetic Data	Moderator: Diego Kaminker, HL7 Panelists: • Farhana Bandukwala, Google • Nikolai Ryzhikov, Health Samurai • Jared Erwin, Microsoft
Commercial Synthetic Data. Is it Coming?	Moderator: Viet Nguyen, HL7 Panelists: • Christopher Hazard, Diveplane Corporation • Noa Zamstein, MDClone
The Use of Synthetic Data for AI/Machine Learning in Healthcare	Moderator: Daniel Vreeman, HL7 Panelists: • Sam Schifman, Availity • Angelo Kastroulis, Ballista Technology Group • Alex Watson, Gretel.ai

Key Themes

Barriers and Limitations of Current Tools

- Synthea is a very powerful tool that can support a variety of use cases for synthetic health data. It was built to meet health IT community demand for electronic health record (EHR) data sets and sidestep cost, privacy, and access issues associated with using de-identified data.
- Synthea is not perfect; like all synthetic generation tools, it has barriers and limitations. Modeling and simulation can be realistic, but users must ask if a synthetic model or the resulting data is useful for a specific use case and validate their findings. Synthea has limitations in statistical accuracy and suitability for deep learning.
- MIMIC IV is a freely accessible, broadly useful database that sources relevant critical care data from an EHR. The project seeks to preserve the contents and terminology of MIMIC IV during conversion to FHIR.
- FHIR is a generalizable and useful tool and provides a good archival model. However, certain types of data are more difficult to transform to FHIR, and the translation of data into FHIR requires additional storage capability--FHIR can be somewhat “verbose” in terms of data sizes.
- Balancing data constraints with coverage (complexity) and tailoring the creation and use of a data set must be considered in the context of the source data or the model being trained. Validation is critical.
- Synthetic Healthcare Database for Research (SyH-DR) is an all-payer, nationally representative claims database that replicates the structure and statistical properties of original eligibility and claims data from Medicare, Medicaid, and commercial health insurance while protecting the privacy and confidentiality of people and institutions. SyH-DR healthcare data supports innovation and evidence-based decision making, provides benchmarking capabilities for healthcare providers, and enables purchasers to understand cost drivers.
- Every data set has pros and cons, and no one data set is appropriate for solving all problems or modeling from all perspectives. Determining whether a data set is fit for purpose requires that users consider whether a data set fulfills their use-case needs and is appropriate to what they are trying to do with the data.
- De-identification of real data is extremely challenging. Users must also determine whether data derived from some “real world” data sets is safe to release publicly.
- Synthetic data is often unable to replicate the “missing elements” in real EHR data, such as inactive conditions and inaccurate data.

Synthetic Data for Testing

- Variability in use cases makes universal solutions for test data sets challenging, and tailoring the creation and use of a test data set must be considered in the context of the source data or the model being trained.
- Input from subject matter experts is essential to creating good test data; it takes a wide range of technical and clinical expertise to support testing needs.
- Synthea is widely used to create test data sets to validate standards, and the diversity and use of Synthea-generated data can be enhanced with new or additional inputs, thus adding value to the larger open-source community.

- Overall, data produced by Synthea is realistic, and the tool is capable of generating large sets of test data. However, Synthea data can be redundant and may require post-processing to trim data from a cohort.

Fairness in Synthetic Data

- When synthetic data is generated from a real data set, the goal is preserving individual privacy while maintaining the properties of the real data. However, when synthetic data is generated, there may be dissimilarity between the real and the synthetic data in terms of subgroups. Avoiding representational bias requires ensuring that the synthetic data generated by a model retains characteristics and proportions similar to the real data.
- The presence of biases within data has proved to be a significant problem in applying AI techniques. Machine learning models can help reduce bias in synthetic data sets, but can also inject new biases; some may be present in the machine learning algorithms.
- Applying a synthetic data generator on data that underrepresents groups of patients can lead to the generation of synthetic data in which specific cohorts of patients may be underrepresented and may lead to structurally missing data or incorrect correlations and distributions.
- There are a variety of techniques for understanding and correcting bias and different approaches for constructing discrimination-free models, including sampling, generative adversarial networks (GANs), and fairness-aware techniques like FairGAN and DeCAF. The effectiveness of techniques for correcting bias varies by use case.
- Currently, no standard set of metrics exists for understanding fairness or defining “good” synthetic data.

The Use of Synthetic Data in Education

- Patient health data contains valuable information and provides a deep-learning test pool for students, but accessing such data is problematic because of patient privacy issues.
- Synthetic data enables a portfolio of educational opportunities accessible to a diverse population of learners. The methods and tools practiced and learned on synthetic data can be readily applied to real world data.
- The applications of synthetic data in educational contexts can help build clinical informatics capacity generally and across a broad range of learners.
- Research shows that learning models trained on computationally derived synthetic data can be as valid as models trained on the original source data from which the synthetic data was derived.
- Through synthetic data, a workforce can explore advances in information and evaluate how the data change over time, and educators can respond with more agility to changing trends in the data by teaching the workforce how to adapt and respond to these changing trends.

A Trove of Synthetic Data, But No FHIR Access

- Synthetic data can be used to modernize health technology systems and improve application quality by providing faster development cycles without the need to access private patient information directly.
- Synthetic data is effective for simple proof of concepts and initial testing of clinical applications, and can be used to ensure that customer and sponsor needs are being met.

- In a FHIR environment, synthetic data can be used to validate clinical definitions.
- Custom modeling with synthetic data allows for testing rare disease scenarios, specific interactions that are difficult to find in real-world datasets, and non-EHR contexts.
- Synthea is quick and effective for generating large numbers of records in FHIR format for import and use. However, simulating interactions across existing modules is difficult. Additional tooling makes Synthea modules comparatively easy to review.
- Functional programming, used to condense and simplify code, offers features for data transformations. Users should consider their use case and goals before initiating a FHIR mapping exercise.

Managing "Big" Synthetic Data

- Large quantities of realistic distributed data are required to test the performance of big data.
- Realistic synthetic data should mirror operational realities (including volume, velocity, and variety), be similar in structure and semantics, and preserve privacy. It requires multiple solutions to achieve these requirements.
- Machine learning and artificial intelligence (AI) techniques can help generate data at scale.
- A primary challenge with large scale synthetic data sets is the time required to generate and load the data. There is a need for faster data generators, staging areas for testing, and tools that can analyze existing data and generate similar production-like data without security risks.
- Validating data platforms requires unit tests, database queries, and sandbox instances.
- Generative testing can help identify issues and ensure comprehensive test coverage.

Commercial Synthetic Data. Is it Coming?

- Machine learning models require large amounts of data with contemporary and relevant content. Unfortunately, clinical data is siloed due to privacy restrictions, and access is often limited to treating clinicians. A methodology is needed to harness clinical data while maintaining patient privacy.
- Differentially private synthetic data sets with mathematical guarantees are promising options for expanding data sharing in ways that have not been possible to date.
- Categorical and date data are often more identifying (creating risk to privacy) than numerical data.
- Some commercial products offer healthcare professionals access to data via query portals to understand their own patient populations.
- When synthetic data becomes acceptable as non-human subject data, these large synthetic data sets can be released without concerns for privacy, thus increasing accessibility and promoting research.

The Use of Synthetic Data for AI/Machine Learning

- Today's healthcare system presents a very fragmented vision of care and the care journey. The volume and complexity of real data derived from patient encounters brings vast computational complexity to solving problems in healthcare, especially because this data complexity is multiplied across domains.

- Real data is messy. Synthetic data has great potential for training and validating AI and machine learning systems; however, because synthetic data is not messy, its use can create significant problems in understanding and solving healthcare problems.
- Some synthetic data sets are enhanced based on statistical models that add structural flaws often observed in real world data. Decision support has been measurably enhanced when comparing dirty synthetic data results with clean synthetic data results.
- When building out synthetic data models that show disease progression, researchers should identify unintended errors in data that can make results difficult to work with.
- Large Language Models (LLMs) hold significant promise for querying patient medical records or healthcare data, but they have flaws. Privacy must be built into the data that LLMs interact with. The alignment and trustworthiness of LLMs must be researched, and knowledge updates and retrieval training are computationally expensive processes.
- AI can deal with unstructured data, but computation rules/clinical quality language (CQL) requires structured data. Both structured and unstructured data are essential.
- AI encodes knowledge by finding patterns in data. Clinical rules are manifestations of clinical guidelines. Creating a framework for generating models influenced by clinical guidelines and produced by experts allows for more explainable models.
- Standards are crucial, even in an AI/Machine Learning environment. Standards provide context for healthcare data in support of semantic interoperability and help lower the bandwidth for AI/Machine Learning models.
- AI is here to stay, and we should think about how to responsibly represent and use it in standards (e.g., as data comes out of AI, how does it appear, how do we understand the provenance, how do we make it useful, etc.). We need to be sure that AI models are tested and comply with standards.
- There is a need and desire for case studies and pilot projects that research synthetic data; once the studies/pilots are complete, the synthetic data should be made openly accessible for the industry.

Future Needs and Recommendations

- Further exploration, validation, and improvement are needed in the areas of data integration, synthetic data, harmonization with standards, and future initiatives. Addressing these items will enhance the usability, quality, and impact of the data provided by Synthea, MIMIC, and other healthcare research databases.
- There is a need for tailoring Synthea data for different output formats and finding strategies for integrating external data sources into Synthea-generated data sets.
- Further exploration is needed on the challenges and limitations of synthetic data, particularly in terms of grammar and complexity. Delving deeper into these challenges and limitations will lead to a better understanding of how synthetic data can be improved.
- More clinical informaticists are needed to help bridge the gaps and build a community of experts who can advance the tools and build/use synthetic data sets.
- The community needs methods for validating structure and semantics of synthetic data sets.
- The healthcare industry, especially large data vendors, must be educated about the availability, promise, and pitfalls of synthetic data.
- A library of training data sets should be established.

- Use case specific, easily accessible, clinically realistic synthetic data sets need to be developed.
- Expanded training and education opportunities for non-technical learners about the availability and use of synthetic data would be helpful.
- Collaborative efforts combining synthetic data from multiple sources for research and other important pilot projects should be supported.
- There is a need for faster data generators, staging areas for testing, and tools that can analyze existing data and generate similar production-like data without security risks.
- Strategies must be explored for making AI/ML useful and available for generating synthetic data sets.
- Responsible representation of AI in standards and ensuring that data produced by AI models aligns with expectations should be supported.
- More industry pilot projects making synthetic data available and usable to the community need to be provided.
- Consider creating standards for making data de-identifiable.

Post Event Analysis

Speakers and participants were invited to complete a post-Event evaluation via *Survey Monkey*. Responses will be used to help strengthen HL7's educational outreach and support its continuing efforts to deliver targeted, beneficial content.

Feedback on the Event was very positive overall. Most participants expressed satisfaction with the information about Synthea. More than 80 percent agreed that they would participate in future events aimed at creating openly accessible synthetic data. Respondents reacted positively to the many perspectives heard from during the Event, appreciated the opportunity for learning from experts about their work, and appreciated hearing about contemporary challenges and solutions. *"I loved that various industry experts shared their work and thoughts, allowing me to learn about the current progress and state of the art."*

When asked which part of the Event was of greatest value to them, responses included: hearing and learning from different voices and perspectives; learning about contemporary approaches, tools, innovations, and limitations; discussions about utility vs. privacy and fairness of synthetic data; Synthea information; and the many use cases for synthetic data. *"Hearing what other companies are doing for synthetic data. It was a great overview of the state of the art and helped solidify our decision to invest in Synthea."*

Participants were asked what "next steps" they identified for themselves based on their Event participation. Among the responses were: reaching out into the user community to bridge and participate in collaborative efforts; sharing materials and take-aways within my organization; exploring Synthea; reaching out to presenters; and investigating and applying Event topics for own work, including creating content. *"[I will] test out the presented solutions, keeping privacy, fairness and utility in mind."*

When asked what HL7 and/or federal agencies can do to drive future development of synthetic test data, a broad range of suggestions were made. One respondent asked that HL7 and/or federal agencies continue educating and bringing in those doing the work. Several respondents focused on areas for standardization, including privacy and utility metrics, education, collaborative initiatives, releasing of public data sets, and standard data sets for use by the health industry. The need for more pilot projects was mentioned. One respondent suggested having individuals with experience in clinical care and workflows and data scientists with systems-level knowledge of healthcare as well as informatics training and experience be included in efforts to develop and identify useful data. Another suggested that HL7 bring together a small, short-term group to help better define distinct data sets, the purposes it can serve, the attributes of the data, and relationship to HL7 standards/FHIR. One respondent suggested partnerships with commercial entities; another suggested forming a coalition of Synthea users to define the industry needs and "kick start" funding, since Synthea funding will soon run out.

"Where [does] HL7 want to go in the future? What is the vision for how synthetic data [and] tools (new or existing) will fit into the FHIR Foundry Concept?"

Participants were asked what synthetic data generator(s) and/or data set(s) they were currently using. More than half the respondents use Synthea and/or internally developed tools/data sets. A small number of respondents use MIMIC, and none reported using SyH-DR. Approximately 15 percent use

commercial products, and more than 25 percent of respondents use other generators and/or data sets.

When asked about their use or planned use of synthetic data, software product testing, implementation guide/standards, and quality measure testing were most often chosen. About 35 percent of respondents reported using synthetic data for AI/machine learning and/or research; almost 20 percent of respondents mentioned education.

Suggestions for Event changes or improvements elicited a broad range of responses, including praise for the Event's good back-and-forth, its relevance and content, and the overall organization of speakers and content. Respondents suggested expansion of contents to include genomics data, privacy metrics and validation, and challenges in realistic use cases. One respondent felt that the virtual environment negatively impacted side conversations and suggested scheduling time for breakout discussions to facilitate additional exchange. An international participant, who relies heavily on the Event recording because of time-zone differences, complained about the recording stopping during some presentations. One respondent, noting that presenters were drawn from the business and technology sectors, suggested more speaker guidance and complained that, even with his technical background, he started to tune out when speakers were talking about minutiae of their implementations. Respondents had different perspectives, based on their knowledge background:

"This seminar was way over my head. I learned a lot but started with zero knowledge of this. It was very interesting seeing all the applications so far and possible future application of synthetic data! I had no idea!"

"There should be a framing overview of the relationship of presentations; no presentations which have no relationship to actual healthcare delivery and clinical practice; and examples for each presentation of how the synthetic data generated would be used relative to informatics, FHIR, and HIT."

"Excellent workshop. All speakers and content were relevant and organized very well. Just the right amount of time."

Acknowledgements

The HL7 FHIR Synthetic Data Event was convened by HL7 International and supported by the Office of the National Coordinator for Health Information Technology (ONC) of the U.S. Department of Health and Human Services (HHS) under grant number 90AX0035 (Title: Closing the Gap Between Standards Development and Implementation).

HL7 would like to recognize the important contributions made by the Event's speakers and the following individuals, whose expertise and guidance also supported this event:

- Patricia Guerra, HL7, Marketing Director
- Dave Hamill, HL7, Program Management Office Director
- Crystal Kallem, C K Consulting LLC, Event Organizer
- Diego Kaminker, HL7, Chief Standards Implementation Officer
- Alex Kontur, ONC, Public Health Analyst
- Laura Mitter, HL7, Design Director
- Viet Nguyen, HL7, Chief Standards Implementation Officer
- Rebecca Parsons, HL7, Senior Program Manager
- Melinda Stewart, HL7, Education Marketing Manager
- Daniel Vreeman, HL7, Chief Standards Development Officer