Synthetic Health Data Challenge Frequently Asked Questions

BASIC FACTS ABOUT PCOR AND SYNTHETIC HEALTH DATA

What is PCOR?

PCOR (Patient-Centered Outcomes Research) is a research field focused on producing scientific evidence comparing the effectiveness of various medical prevention and treatment options while also considering patients' health care preferences, values, and the questions they face when making health-care decisions. Robust data infrastructures that support rigorous analyses and generate relevant information strengthen the validity of PCOR findings.

What are synthetic health data?

Synthetic health data (sometimes called synthetic health records) is realistic (but not real) patient data and associated health records. This realistic data for fictional patients, which models patients from birth until death, is free of protected health information (PHI) and personally identifiable information (PII) constraints. Synthetic health data can be generated to meet the specific interests of PCOR researchers and developers for testing theories, data models, algorithms, and prototype innovations.

Why is synthetic health data important?

Researchers and developers depend on clinical data for testing research algorithms and/or technology while awaiting access to real clinical data. Unfortunately, cost, patient-privacy concerns, and other legal restrictions can make high quality, health- and health-care related data difficult to access. Anonymized data (data from the health records of actual patients with personal information stripped away) is often used. However, the risk of re-identification of anonymized data is high and, especially for rare conditions, impossible to completely eliminate.

Further, because of a variety of interoperability issues, it can be difficult to bring data together from different resources for the purpose of robustly testing analysis models, algorithms, or assisting in the development of software applications. After securing data, there are several processes that must be done before beginning to apply or use the data. For example, a researcher or heath IT developer will typically need to aggregate, de-identify, and analyze data before testing the effectiveness of algorithms and modeling approaches used in matching and disease modeling techniques. Interoperability issues also make it difficult to compile large amounts of data from different sources for the purposes of robustly testing analysis models or assisting with the development of software applications. Synthetic health data also offers the kind of built-in interoperability and integration of clinical and claims data that rarely exists in the real world.

BACKGROUND ON PCORTF, ONC, AND FEDERAL SUPPORT FOR CHALLENGES

What is PCORTF?

The Office of the Secretary (OS) of Health and Human Services, through the Patient-Centered Outcomes Research Trust Fund (PCORTF), is charged with coordinating relevant federal health programs to build data capacity for comparative clinical effectiveness research. Information about the OS-PCORTF Strategic Framework for PCOR data and its use can be found at https://aspe.hhs.gov/patient-centered-outcomes-research-trust-fund-faqs.

Who is ONC?

ONC (Office of the National Coordinator for Health Information Technology) is a federal entity that, as part of its broad public-service mission to coordinate nationwide efforts to implement and use the most advanced health information technology and the electronic exchange of health information, leads and collaborates on projects that inform policy, standards, and services specific to the adoption and implementation of a data infrastructure for PCOR. One of ONC's current projects, funded by PCORTF, is the Synthetic Health Data Generation Engine to Accelerate PCOR. The Synthetic Health Data Challenge is an important component of that project. More information about ONC can be found at https://www.healthit.gov/.

Why does the Federal government support Challenges?

Challenges enable the Federal government to tap into the expertise and creativity of the public. Under a directive calling for innovative ways to generate ideas and collaboration, Challenges are policy tools that can foster participation in government activities through the process of co-creation. Challenges may offer a variety of prizes, including cash, recognition, or the deployment of a winning solution. For more information about Challenges, visit https://www.healthit.gov/topic/innovation/health-it-prizes-and-challenges-faqs.

SYNTHETIC HEALTH DATA CHALLENGE OVERVIEW

What were the goals of the project that conducted the Challenge?

The Synthetic Health Data Challenge is an important part of the Synthetic Health Data Generation Engine to Accelerate PCOR project. The goal of the project is to enhance Synthea™, an open-source synthetic health-data generator, and to support PCOR research needs by increasing the number and diversity of available synthetic patient records. The project targets the areas of opioids, pediatrics, and complex care, because of the unique characteristics of these data needs. Increased availability of synthetic data for these priority areas will help expedite testing of research algorithms and technology. Part of the project evaluated existing Synthea data modules to assess opportunities for development and enhancement.

What was the Synthetic Health Data Challenge?

The Synthetic Health Data Challenge (Challenge) is an important part of the Synthetic Health Data Generation Engine to Accelerate PCOR project. The Challenge was a prize competition that invited innovators, researchers, and technology developers to create and test innovative and novel solutions aimed at further cultivating the capabilities of Synthea and the synthetic health data it generates. The Challenge took place between January 19, 2021 and July 13, 2021, and was implemented to demonstrate novel uses and validate the realism of Synthea-generated synthetic health records. Submissions were evaluated by a panel of judges and the Challenge awarded $100,000 in total prizes, including one First Place ($40,000),

two Second Places ($15,000), and three Third Places ($10,000). Each winning entry presented their work during the Synthetic Health Data Challenge Winning Solutions Webinar on October 19, 2021.

Why did the Challenge focus on enhancing Synthea and Synthea-generated data?

Clinical data are critical for the conduct of PCOR, which focuses on the effectiveness of prevention and treatment options. However, realistic patient data are often difficult to access because of cost, patient privacy concerns, or other legal restrictions. Synthetic health data can help address these issues and speed the initiation, refinement, and testing of innovative health and research approaches. ONC conducted the Challenge to enhance Synthea to accelerate research and support the greater PCOR data infrastructure and capacity by providing researchers and health IT developers with a low-risk, readily available synthetic data source to provide access to data until real clinical data are available. To learn more about Synthea, visit: https://github.com/synthetichealth/synthea/wiki

What Synthea modules were Challenge participants required to use?

Participants could use any published Synthea module to support their proposed solution as long as the proposal related to one of the three PCOR priority use cases (opioids, complex care, and/or pediatrics).

ABOUT SYNTHEA™

What is Synthea? What are clinical disease modules?

Synthea (pronounced [SIN] + [THEE] + [UH]) is an open-source, fully synthetic set of electronic health record data developed by the MITRE Corporation that can be used to model a vast array of disease states and populations. Detailed information for using Synthea can be found here: https://github.com/synthetichealth/synthea/wiki.

Synthea's clinical disease modules are created by the community using a combination of clinical care protocols, publicly available disease incidence and prevalence statistics, and clinical expert feedback. Synthea uses these modules to generate synthetic health records, simulating the progression and treatment of disease from birth to death. Currently there are more than 120 modules and submodules available in Synthea which can be found in the Synthea GitHub Repository here: https://github.com/synthetichealth/synthea.

What data sources does Synthea use?

The Synthea software uses a temporal model to generate the medical history of synthetic patients. Synthea contains publicly available demographic data obtained from the United States Bureau of the Census. The data are post-processed to create population input data for locations in the United States. This post-processed data can be used with Synthea to generate representative populations. Synthea's clinical disease modules are often designed using publicly available disease incidence and prevalence statistics and clinical care maps.

What are the known limitations of Synthea?

Examples of known limitations include:

- Synthea modules are often built using clinical care guidelines and standards of care. As a result, the data generated does not contain variations in care which would occur in the real world.
- Synthea data focus solely on care provided in the hospital and provider settings. Behavioral therapies and treatments administered outside of the hospital are not included.
- Replicating population-level statistics using Synthea is challenging because each run is different due to the random nature of the simulation.

Additional information about Synthea and synthetic data limitations is discussed in the following publication: *Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association. 2018 Mar 1;25(3):230-8. Available at https://pubmed.ncbi.nlm.nih.gov/29025144/.*

What standard terminology systems are used by Synthea?

Synthea relies upon SNOMED-CT and LOINC terminology systems, which are freely available to the community.

In what standard formats are Synthea's synthetic health records generated?

Synthetic health records are generated in a variety of formats, including text, HL7 FHIR®, comma-separated values (CSV), and HL7 C-CDA®