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.

OVERVIEW OF THIS CHALLENGE

Is the Challenge part of a larger project?

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. To learn more about the Synthetic Health Data Generation Engine to Accelerate PCOR project, visit https://www.healthit.gov/topic/research-evaluation/synthetic-health-data-generation-accelerate-patient-centered-outcomes. To learn more about Synthea, visit: https://github.com/synthetichealth/synthea/wiki

What is the Synthetic Health Data Challenge?

The Synthetic Health Data Challenge (Challenge) is a prize competition that invites 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 was implemented to

demonstrate novel uses and validate the realism of Synthea-generated synthetic health records. Information about the Challenge is available at: https://www.challenge.gov/challenge/synthetic-health-data-challenge

Must I be a U.S. citizen to participate in the Challenge? *02/09/21*

To participate in the Challenge, you must be a citizen or a lawful permanent resident (i.e., green card holder) of the U.S. Individuals holding non-resident student visas (e.g., an F-1 vista) or work visas (e.g., EB-1, -2, or -3 visas) are not eligible to participate in the Challenge, even if they are employed by a U.S. corporate, non-profit, or academic institution. Eligibility requirements are located under ***Rules*** at: https://www.challenge.gov/challenge/synthetic-health-data-challenge

Can teams of students or others participate in the Challenge? *02/09/21*

The Challenge welcomes team participation. No limitations are placed on team membership as long as each team member meets the Challenge eligibility requirements. Each team must designate a team leader, who serves as the primary point of contact for the Challenge. Information about the team leader and all team members must be provided on the Registration Form that is required as part of all Phase I submissions. The Synthetic Health Data Challenge Registration Form is located under ***Resources*** at: https://www.challenge.gov/challenge/synthetic-health-data-challenge

What are the categories for Challenge entries?

The Challenge has two entry categories:

Category I – Enhancements to Synthea. Solutions in this category include, but are not limited to, development or enhancement of Synthea modules and development of solutions that enhance or address limitations of Synthea.

Category II – Novel Uses of Synthea-Generated Synthetic Data. Solutions in this category include, but are not limited to, novel uses of Synthea generated data for research and technology development.

When does the Challenge take place?

The entire Challenge takes place between January 19, 2021 and July 13, 2021.

What are the Challenge phases?

The Challenge is broken into two (2) phases:

Phase I – Proposal for Innovative Models: Participants will submit a written proposal describing their proposed solution, including methodology and intended outcomes. Selected Phase I proposals will proceed to Phase II.

- Proposals are invited from teams or individuals.
- Any number of proposals can be submitted.
- Judges use scoring criteria to determine the Phase I proposals that move on to Phase II.

Phase II – Prototype/Solution Development: Phase I proposals that are selected to proceed to Phase II will develop their prototype/solution at this stage.

- Only Phase I winning proposals compete during Phase II.
- Judges use scoring criteria to determine the winning solutions from among the Phase II submissions.
- Winning solutions are eligible for cash awards.

How many participants/teams are expected to reach Phase II? *02/09/21*

The number of Phase II participants will be determined by the number of Phase I proposals that meet the requirements for moving on to Phase II. A judging panel is tasked with evaluating Phase I proposals and selecting those that will move to Phase II.

What awards are Challenge participants eligible for?

The Challenge offers up to $100,000 in total prizes:

- Up to two (2) first place winners will receive $25,000.
- Up to two (2) second place winners will receive $15,000.
- Up to two (2) third place winners will receive $10,000.
- Honorable Mentions may be awarded but will not receive a monetary prize.

As an additional incentive, each winning entry and honorable mention will be invited to present during a public webinar in October 2021.

How will entries be judged? *02/09/21*

Technical reviewers with expertise relevant to the Challenge will evaluate Phase I and Phase II submissions based on a list of criteria detailed on the Challenge webpage. The submissions and evaluation statements will then be reviewed by federal employees serving as judges. Judges have, at a minimum, basic knowledge and understanding of synthetic health data, PCOR research, and health IT standards and tools.

For Phase I, the judges will determine which submissions will move on to Phase II. For Phase II, the judges will select up to six (6) Challenge winners and may also name honorable mentions. Judging is subject to a final decision by the Award Approving Official: Micky Tripathi, PhD, National Coordinator for Health Information Technology.

Must submissions address one of the example Synthea modules developed by the Synthetic Health Data Team (cerebral palsy, sepsis, spina bifida, and opioid use)? *02/09/21*

Participants are not required to use one of the four example modules (cerebral palsy, sepsis, spina bifida, or opioid use). Participants may use any Synthea module they choose to support their proposed solution as long as the proposal relates to one of the three PCOR priority use cases (opioids, complex care, and/or pediatrics).

Is the Challenge seeking any specific outcomes for the priority use cases (opioids, complex care, and/or pediatrics)? *02/09/21*

The Challenge is not looking for specific outcomes within any of the priority use cases. The possibilities are deliberately broad to fully engage the expertise and innovative skills of each participant.

Where is there more detailed information about the Challenge?

For more detailed information about all aspects of the Challenge, including how to enter, eligibility rules, key dates, and submission requirements, go to https://www.challenge.gov/challenge/synthetic-health-data-challenge.

Who should be contacted with questions about the Challenge?

Send feedback or questions about the Challenge to SyntheticDataChallenge@govhealth.com.

ABOUT SYNTHEA™

How is Synthea pronounced? *04/15/21*

Synthea is pronounced [SIN] + [THEE] + [UH], just like the name, Cynthia.

What data sources does Synthea use? *02/09/21*

The Synthea software uses a temporal model to generate the medical history of synthetic patients. Synthea contains publicly available demographic data obtained from the U.S. Census Bureau. 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? *02/09/21*

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? *02/23/21*

Synthea relies upon SNOMED-CT and LOINC terminology systems, which are freely available to the community. The Synthea Wiki describes which code systems are allowed for each state at https://github.com/synthetichealth/synthea/wiki/Generic-Module-Framework%3A-States. If a code is allowed for a given state, it will be specified under the Supported Properties / code attribute.

How are Synthea-generated data validated? *02/09/21*

Validation is a requirement for all Challenge submissions. How validation is accomplished is based on module design and the validation strategy defined by each participant.

Are participants limited to using or enhancing Synthea or its data? Can participants create solutions using their own observational or clinical research data? *02/09/21*

The Challenge requires that participants use Synthea and Synthea-generated data. However, this requirement does not preclude the use of a participant's source data as part of modeling and/or data validation processes. For example, participants may want to compare Synthea output to their own data sources as part of a validation component.

What are the goals for enhancing Synthea and Synthea-generated data? *02/09/21*

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. As such, ONC is issuing this challenge in an effort 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.

PHASE II SOLUTIONS: GUIDANCE FOR PARTICIPANTS

How are placements and award amounts determined? *04/15/21*

The Challenge has three award categories and up to two winners will place in each category. First place winner(s) will receive $25,000 -- $50,000; second place winner(s) will receive $15,000 -- $30,000; and third place winner(s) will receive $10,000 -- $20,000. The number of first, second, and third place winners as well as the amounts awarded to each winning solution will be determined by the number and quality of the Phase II submissions and the judges' scoring of each submission.

What should be submitted to the GitHub repository? *04/15/21*

All non-proprietary source code developed as part of the solution should be uploaded to GitHub. A URL to the source code in GitHub should be included in the final paper.

Are more points awarded in the implementation category if the solution is written in Java vs. other programming languages? *04/15/21*

The Challenge does not require participants to use a specific programming language. If the solution would benefit the Synthea community, participants are encouraged to align solution development with guidance provided on the Synthea Wiki related to Contributing.

Must a solution be easily replicated/used by non-technical clinicians or software developers? *04/15/21*

The extent to which other developers and/or implementers can reproduce and/or reuse the solution will be evaluated and scored by the judges. However easy replication/use by non-technical clinicians is not expected.

What does good validation look like? What findings are expected as a result of the validation process? *04/15/21*

Validation depends on the requirements of the specific solution, so can only be determined by the participant. The validation should demonstrate how a solution supports and/or improves the validation capabilities of Synthea, improves the clinical relevance of the Synthea-generated synthetic data, and/or supports others who may want to use or reproduce the validation method. Participants may include any kind of validation testing that they determine would benefit their solution.

Will this Challenge be held again? *04/15/21*

This Challenge is the first one of its kind targeting the use case areas of pediatrics, complex care, and opioids. The results of this Challenge will give ONC insight into the benefits that this medium offers for tackling tough problems. The ONC News and Updates is a good source of information about future Challenges, PCOR-related work, and other opportunities.