The Office of the National Coordinator for Health Information Technology

Synthetic Health Data Generation to Accelerate Patient-Centered Outcomes Research (PCOR)

FINAL REPORT

PREPARED BY

Clinovations Government + Health for the
Office of the National Coordinator for Health Information Technology

Contract No.: HHSP233201500099I/75P00119R00375

March 2022

Table of Contents

Acknowledgements

The authors of this document are:

- Stephanie Garcia, MPH, Office of the National Coordinator for Health Information Technology (ONC)
- Crystal Kallem, RHIA, CPHQ, Clinovations Government + Health
- Casey Thompson, MSN, RN-BC, Clinovations Government + Health
- Maureen Tan, Clinovations Government + Health
- Yan Heras, PhD, Optimum eHealth, LLC
- Daniella Meeker, PhD, Keck School of Medicine, University of Southern California

The authors would like to recognize the important contributions made by the following individuals who shared their expertise and guidance throughout the project:

- Brittney Boakye, MPH, ONC
- Kevin Chaney, MGS, Department of Veteran Affairs
- Bo Dagnall, MS, Perspecta
- Laurie Glader, MD, Nationwide Children's Hospital
- James Hellewell, MD, MS, Intermountain Healthcare
- Jennifer Judy, PhD, Pfizer (formerly with The National Institutes of Health (NIH))
- Thomas Kannampallil, PhD, Washington University School of Medicine, St. Louis
- William (Bill) Lawrence, MD, Patient-Centered Outcomes Research Institute (PCORI)
- Ted Melnick, MD, MHS, Yale School of Medicine
- Lisa Mirel, MS, Centers for Disease Control and Prevention (CDC)
- Penny Mohr, MA, PCORI
- Stuart Myerburg, JD, CDC
- Viet Nguyen, MD, Stratametrics
- Pradeep Podila, PhD, CDC
- Anita Samarth, Clinovations Government + Health
- Carmen Smiley, PhD, ONC
- Jason Walonoski, MS, The MITRE Corporation
- Li Xiong, PhD, Emory University
- Alda Yuan, JD, ONC
- Teresa Zayas Cabán, PhD, NIH

Executive Summary

The *Synthetic Health Data Generation to Accelerate PCOR* project[1] (project) was launched in 2019 by the Office of the National Coordinator for Health Information Technology (ONC). This project is part of ONC's portfolio of patient-centered outcomes research (PCOR) projects funded by the PCOR Trust Fund that is administered by the Department of Health and Human Services (HHS) Assistant Secretary for Planning and Evaluation (ASPE). ONC is the principal federal entity charged with coordination of nationwide efforts to implement and use the most advanced health IT and the electronic exchange of health information; the agency serves as a resource to the entire health system to support adoption of health IT and promotion of nationwide, standards-based health information exchange.[2] The project sought to enhance the ability of Synthea™, a synthetic health data generator,[3] to produce high quality synthetic health data and increase the number and variety of available synthetic health records. With specific focus on complementing the PCOR data infrastructure, the project was informed by ONC's goals of fostering research, scientific knowledge, and innovation, and enhancing the nation's health IT infrastructure.[2]

PCOR uses scientific evidence to compare the effectiveness of various prevention and treatment options and care delivery models while also considering patients' health care preferences, values, and the questions they face when making health-care decisions.[4] Clinical health data are critical for conducting PCOR. However, high quality, real clinical health data can be difficult to access because of complex privacy concerns, security restrictions, and usage issues. As a result, PCOR researchers, health information technology (health IT) developers, and informaticians often depend on anonymized or de-identified clinical health data for testing their theories, data models, algorithms, and prototype innovations. However, re-identification of anonymized data remains a possible security risk.

Synthetic health data can provide a lower risk data source to complement research use of real clinical health data and meet testing needs until real clinical health data are available. A synthetic health data generation engine can generate realistic synthetic health data reflecting the characteristics of a population. Synthetic health data may also offer built-in standardization of clinical and claims data that rarely exists in the real world. Additionally, synthetic health data can be shared among teams and with external partners. Access to synthetic health data while awaiting access to real clinical data may enhance researchers' ability to effectively conduct rigorous analyses and produce relevant findings to inform health and treatment decisions.[4] The availability of reliable and robust synthetic health data generation engines can also safeguard the privacy of real patients because it supports appropriate stewardship practices in which real clinical health data are only accessed and used when necessary.

Synthea generates realistic, but not real, health data for fictitious patients. Synthea modules serve as the foundational element for the creation of this synthetic health data by capturing the parameters for conditions, encounters, laboratory tests, and other clinical and demographic information. The modules are used by the Synthea engine to generate data for a given patient population and geographic location. Synthea-generated synthetic health data are conformant with many data formats, and Synthea's FHIR data output also supports the United States Core Data for Interoperability (USCDI), a standardized set of health data classes and constituent data elements that enable nationwide, interoperable health information

exchange.[5] The project developed and published five new Synthea modules directed at increasing the number and diversity of Synthea-generated synthetic health records available for PCOR use cases. The focus areas for module development—patients with complex care needs, opioid use, and pediatric populations—were selected because these use cases are associated with a higher likelihood of re-identification, real health data that are typically more difficult to access, and additional privacy considerations that may not impact real clinical health data from other use cases. The importance of these areas to ONC is evidenced by prior and ongoing work related to pediatric populations[6] and opioids.[7] The project's five modules are titled *Prescribing Opioids for Chronic Pain and Treatment of Opioid Use Disorder*, *Treatment of Sialorrhea in Cerebral Palsy*, *Sepsis*, *Spina Bifida*, and *Acute Myeloid Leukemia*.

The Synthetic Health Data Challenge (Challenge) competition engaged innovators, researchers, and health IT developers to create and test novel solutions that further cultivated the capabilities of Synthea and the synthetic health data it generates. A team of pharmacists won first place for their *Medication Diversification Tool* that enhances Synthea data sets by increasing the variation of medication orders in Synthea's Generic Module Framework (GMF).[8] Finally, a demonstration study assessed whether Synthea might serve as a tool for PCOR hypothesis testing. The Acute Myeloid Leukemia (AML) module was designed to replicate a simulation comparing levofloxacin prophylaxis to usual care for leukemia patients undergoing chemotherapy. By comparing Synthea's GMF to platforms designed for simulation, the study demonstrated that the Synthea platform can be used, with modifications, for some types of simulation studies.

CONSIDERATIONS FOR ADVANCING SYNTHEA FOR PCOR

Each aspect of the project resulted in findings and lessons learned, with a focus on the development, testing, and validation of new Synthea modules; and enhancing the community's awareness of Synthea and its capabilities. Themes included the importance of engagement by experts with diverse backgrounds and interests; the requirement for a systematic, iterative module development process; and the importance of documentation to support current and future Synthea users. A broad range of opportunities exist for community advancement of Synthea for PCOR. Analysis of the successes and obstacles encountered informed a list of suggestions that would serve to enhance Synthea software, guidance, and documentation. When implemented, these changes will help make Synthea software more capable and user-friendly while proving additional support to encourage both novice and experienced users. The sustainability of Synthea's evolution relies on community participation, so efforts to enhance and maintain community-wide stewardship for Synthea are essential to its overall success and continuing usefulness to PCOR stakeholders.

CONCLUSION

The project met its goal of supporting PCOR for use cases in the focus areas of patients with complex care needs, opioid use, and pediatric populations by enhancing Synthea's ability to produce high quality synthetic health data and increasing the number and variety of Synthea-generated synthetic health records. The project also bolstered ongoing collaboration among the synthetic health data community, clinicians, and researchers, which is essential to outputting more accurate synthetic health data and advancing the use of synthetic health data to accelerate PCOR research and development of health IT. Further, the availability of reliable and robust synthetic data generation tools can safeguard patient privacy because they support appropriate stewardship practices in which real patient data is only accessed and used when necessary.

Introduction

BACKGROUND

The *Synthetic Health Data Generation to Accelerate PCOR* project[1] (project) is one in a portfolio of patient-centered outcomes research (PCOR) projects led by the Office of the National Coordinator for Health Information Technology (ONC). Projects in this portfolio are funded by the PCOR Trust Fund, which is administered by the Department of Health and Human Services (HHS) Assistant Secretary for Planning and Evaluation (ASPE). This body of work informs and contributes to the development of policy, standards, and services that are specific to the implementation of a data infrastructure for the conduct of PCOR. Robust data infrastructure that supports rigorous analyses and generates relevant information strengthens the validity of PCOR findings.

PCOR produces scientific evidence comparing the effectiveness of various prevention and treatment options and care delivery models while also considering patients' health care preferences, values, and the questions they face when making health-care decisions.[4] Real clinical health data (data from the health records of real patients) are critical for PCOR, but high quality, real clinical health data can be difficult to access because of cost, patient privacy concerns, or other legal or institutional review board (IRB) restrictions. PCOR researchers, health information technology (health IT) developers, and informaticians often depend on anonymized or de-identified real clinical health data for testing theories, data models, algorithms, and prototype innovations; but may be required to aggregate, de-identify, or analyze that data before use. However, the risk of re-identification of anonymized real clinical health data is impossible to eliminate, especially for rare medical conditions.[9],[10] Real clinical data are not easily shared among teams and with external (research and developer) partners. Additionally, issues with interoperability can make it difficult for researchers to compile copious amounts of real clinical health data from different sources for the purposes of robustly testing analysis models, algorithms, or developing software applications.

Synthetic health data can complement the use of real clinical health data in research settings. A robust and readily available synthetic health data generation engine can augment the PCOR infrastructure by providing researchers and health IT developers a low-risk source of synthetic health data as they await access to real clinical health data.[4] Synthetic health data can be shared easily and may also offer built-in standardization of clinical and claims data that rarely exists in the real world. The availability of reliable and robust synthetic health data engines can also safeguard the privacy of real patients because it supports appropriate stewardship practices in which real clinical health data are only accessed and used when necessary.

Synthea™ (pronounced sin-thee-uh),[3] is an open-source synthetic health data generator (i.e., a synthetic health data generation engine) created by the MITRE Corporation[11] and used for this project. Synthea can meet researchers' testing needs by generating realistic (but not real) synthetic health data that can reflect the specific characteristics of a population of interest. Synthea models the medical history of synthetic patients[3] from birth until death and is free of protected health information (PHI) and personally identifiable information (PII) constraints because the data are generated randomly and independently from publicly available datasets. For example, Synthea uses United States Bureau of the Census (Census) data to

generate a random distribution of initial demographic-specific conditions reflecting local populations.[12] Synthea data sets are compatible with a variety of technologies, including Health Level Seven International® (HL7®) Fast Healthcare Interoperability Resources® (FHIR®) and Consolidated-Clinical Document Architecture (C-CDA). Synthea-generated synthetic health data and associated health records are conformant with FHIR (R4, STU3, and DSTU2), US Core IG, CCDA, and other formats. US Core descends directly from the Argonaut guide to support FHIR Version STU3 and, eventually, FHIR R4 and the ONC U.S. Core Data for USCDI v1. USCDI is a standardized set of health data classes and constituent data elements for nationwide, interoperable health information exchange.[5] Synthea is used and enhanced by a community of developers, academics, and health care experts. Although Synthea is an open-source tool, it is actively supported by a team of MITRE developers who address new feature requests, consider the clinical validity of contributed modules, and manage other issues identified by the user community.

Synthea modules are used by Synthea to generate synthetic health data and associated synthetic health records covering a vast range of health care scenarios. These synthetic health data are free to use and do not contain the privacy concerns, security restrictions, and usage issues of real clinical health data. Synthea and its synthetic health data can be used for a variety of secondary applications in academia, research, industry, and government.[13]

This project focused on use cases in the HHS and ONC priority areas of patients with complex care needs, opioid use, and pediatric populations. For example, opioid addiction is one of the most critical public health crises facing our nation, and addressing it is one of HHS's top priorities. These focus areas were also selected because the use cases are associated with a higher likelihood of re-identification, the real health data are typically more difficult to access, and there are additional privacy considerations that may not impact real health data from other use cases.

PROJECT GOAL

The project goal was to complement the PCOR infrastructure by expanding the capabilities of Synthea to generate high-quality synthetic health data, thus increasing the quantity and variety of synthetic health records available for researchers, health IT developers, and informaticians. The project achieved this goal by:

- Identifying and convening a multidisciplinary technical expert panel (TEP) to provide insights regarding the selection of use cases and module development;
- Developing Synthea synthetic health data generation modules reflecting use case areas of patients with complex care needs, opioid use, and pediatric populations; and
- Engaging a broader community of researchers, developers, and innovators to validate the realism and demonstrate the potential uses of Synthea-generated synthetic health data through a nationwide challenge competition: the Synthetic Health Data Challenge (Challenge).

This report details project activities, findings, and considerations for enhancing Synthea to support and accelerate PCOR.

Technical Expert Panel (TEP)

A Technical Expert Panel (TEP) was assembled from key stakeholder groups. Targeted areas of expertise encompassed PCOR research, clinical care, epidemiology, FHIR®, informatics, opioids, pediatrics, synthetic health data/Synthea, and technology. The TEP was tasked with:

- Recommending modules to be developed or updated in the focus areas of patients with complex care needs, opioid use, and pediatric populations;
- Recommending publicly available data sources and information that might serve as the basis for module development and testing;
- Providing subject matter input for module design and development; and
- Providing input into the development of a demonstration study as an example of how synthetic health data generated by this project might serve as a tool for PCOR hypothesis testing.

The TEP kick-off meeting was in December 2019, followed by a January 2020 in-person meeting. Virtual meetings followed in February, June, August, and November of 2020 and January, March, and August of 2021. The close-out meeting was in November 2021. The meetings featured formal presentations and facilitated discussions regarding use case selection, progress of module development, and issues encountered during the development process. Some meeting highlights are detailed below:

- December 10, 2019: A brainstorming session provided insight into researchers' synthetic health data requirements, suggested potential module use cases based on the three focus areas, and identified facilitators and barriers to prioritizing and developing potential module use cases.
- January 15, 2020: Based on break-out session discussions and the availability of clinical guidelines, the TEP recommended two opioid-related use cases for development and/or enhancement. Members also suggested evaluating the feasibility of a third use case related to transitioning from acute pain to chronic pain. The two opioid use cases were later merged, resulting in the Prescribing Opioids for Chronic Pain and the Treatment of Opioid Use Disorder (OUD) module. The acute to chronic pain transition was not pursued in this project.
- February 27, 2020: Based on TEP recommendations, registry data elements from the Cerebral Palsy Research Network (CPRN) Cerebral Palsy (CP) registry were evaluated for the development of a CP module. Several of the American Academy for Cerebral Palsy and Developmental Medicine's CP care pathways [14] were discussed, including dystonia, hip surveillance, osteoporosis, central hypotonia, and sialorrhea. The resulting Treatment of Sialorrhea in Cerebral Palsy (CP) module models the treatment of sialorrhea in CP for patients 18 years of age or younger.

- June 4, 2020: The TEP considered sepsis, autism, and spina bifida use cases for development and discussed the feasibility of developing modules focused on diabetes, transitioning from acute to chronic pain, leukemia and pediatric brain cancer, and cervical cancer. Based on the project use case selection criteria (Table 1) and a module use case feasibility assessment that considered the number of related PCORI-funded projects, the TEP recommended sepsis and spina bifida as module candidates. This resulted in a Sepsis module focusing on diagnosis and first hour of care in the ICU and a Spina Bifida module modeling myelomeningocele ("open spina bifida").
- August 25, 2020: Based on feedback from the TEP, the Treatment of OUD submodule was merged into the Prescribing Opioids for Chronic Pain module, enabling the clinical pathway to be triggered by other modules. For the Sepsis module, the TEP suggested adding disposition outcomes, such as Acute Respiratory Distress Syndrome (ARDS), to the module along with expanding the broad-spectrum antibiotic choices and adding vitals and other temporal data. In addition, the TEP recommended adding neonatal surgery to the Spina Bifida module.
- November 4, 2020: A wide-ranging discussion regarding blood pressure vs. mean arterial pressure (MAP) in the Sepsis module led to a recommendation that MAP be added to the module in addition to blood pressure readings. The TEP confirmed mortality rates in the Spina Bifida module.
- Mar 31, 2021: The TEP suggested that, although Synthea cannot create a population of patients who reflect the complexity of the real world, it is a starting point for representing different populations; and that the Demonstration Study appeared to produce data sufficiently valid for the purposes of software development, testing, and demonstration.

TEP members were also project advocates, informally disseminating information to colleagues and helping expand the overall public view of the project. For example, three TEP members served as panelists for presentations at the 2020 American Medical Informatics Association (AMIA) Annual Symposium and the 2021 ONC Annual Meeting.

Module Development

Synthea modules are the foundational element that drives the creation of synthetic health data by capturing the parameters for conditions, encounters, laboratory tests, and other clinical and demographic information. The modules are used by Synthea to generate synthetic health data and associated health records for a given patient population and geographic location. Synthea houses more than 90 modules and submodules covering a range of disease conditions. This brief overview of Synthea technology will support understanding of the module development process.

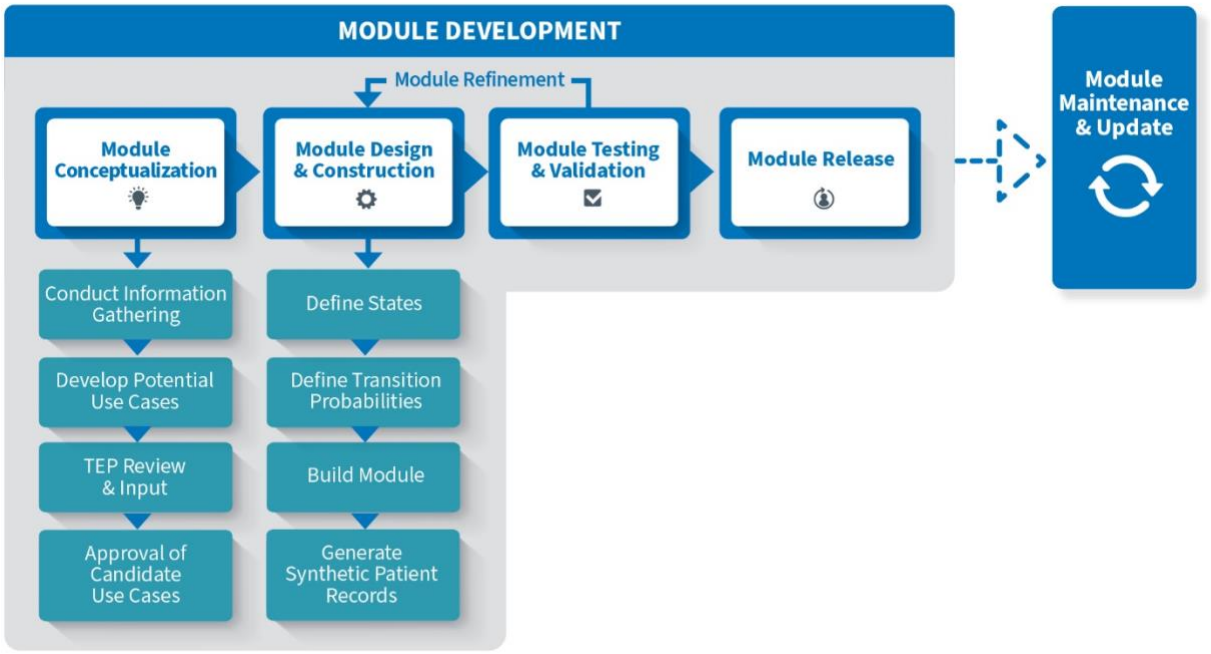
- Synthea modules are written using JavaScript Object Notation (JSON) and built based on publicly available health statistical data as well as other health-related resources, including Centers for Disease Control and Prevention (CDC) incidence and prevalence rates, clinical care pathways and guidelines, standards of care, literature articles, and National Institutes of Health (NIH) reports; these are used to simulate disease conditions where events occur with varying frequency based on repeated, weighted random sampling.[13]
- Synthea provides a GMF for representing disease and clinical treatment models using a set of predefined states, transitions, and conditional logic. States or transitions in a Synthea module can trigger clinical events, such as condition onsets, encounters, and medication prescriptions. Most modules are created using the Synthea GMF.[8] In some instances, additional customization is required, which can lead to increased levels of effort during the module development and validation cycles.
- Synthea enables the creation of main modules and submodules. Main modules represent a particular disease or condition. Submodules are smaller, reusable components that can be called upon at any time as directed by a main module. Submodules may be used once or repeatedly in the same module or by several different modules.
- The Synthea Module Builder[13,15] (module builder) is a web interface that makes it easier for non-developers to design and build modules. It supports the GMF specification and allows users to create full-featured modules without needing to directly edit files in JSON. JSON is the source file format used by Synthea to represent modules. It also allows users to download a newly created or edited module from the module builder as a JSON file and to open a module in the module builder using a locally saved module JSON file. This round-trip capability makes it feasible to develop modules with the module builder and then execute the modules locally, hence supporting the iterative module development process. The module builder includes all modules and submodules contained within the Synthea GitHub Repository.[16] Users can edit those modules or create new ones to extend the capabilities of Synthea.

Synthea modules serve as the foundational element that drives the creation of synthetic health data.

MODULE DEVELOPMENT APPROACH

A module development methodology was created to guide the selection of candidate use cases and the subsequent development of new Synthea modules. The methodology outlined a lifecycle for module development, beginning with conceptualization, moving iteratively through design and construction, testing and validation, and culminating with inclusion in the Synthea GitHub repository[16] for use by the community. Figure 1 provides an overview of the module development methodology. The subsequent sub-sections include descriptions of the methodology and module developer findings as they applied this approach.

Figure 1: Module Development Lifecycle



Module Conceptualization

During module conceptualization, PCOR use cases within each of the three focus areas were identified and prioritized for possible development as Synthea health data generation modules. The evaluation criteria detailed in Table 1 helped guide use-case identification and prioritization.

Table 1: Use Case Evaluation Criteria

Evaluation Criteria	Use Case Characteristics
Importance	• Clinical significance of the use case aligns with the focus areas identified to accelerate PCOR research: patients with complex care needs, opioid use, and pediatric populations. • Use case increases the number and variety of Synthea-generated synthetic health records.

Evaluation Criteria	Use Case Characteristics
Feasibility	• Clinical care map is available or can be readily constructed to support module development. • Incidence and prevalence statistical data are available. • Complexity of use case scenarios, flows, and of clinical concepts can be supported by the Synthea GMF and, therefore, by the module builder.
Reliability	• Evidence supporting the clinical care maps, the incidence rates, and the prevalence rates identified for the use case are scientifically reliable.
Use	• Use case does not require a model for deviations in care or potential outcomes resulting from care deviations.[17] (See Opportunities to Enhance Synthea Software.) • Use case does not require realistic representation of poor data quality captured at the point of care or individual deviations in health or outcomes data. (See Opportunities to Enhance Synthea Software.) • Use case allows for validation of its realism and/or potential use of its generated synthetic health data and associated synthetic health records by Challenge participants and the broader community.
Related/Existing Modules	• Existing modules and submodules were analyzed to determine whether new modules should be developed or existing modules updated.

Module Design and Construction

During module design and construction, the module builder was used to develop modules for the selected use cases. Additional publicly available statistics and clinical care guideline information was gathered, and each module was iteratively developed and refined with increasing degrees of detail and accuracy. Throughout the process, the TEP was consulted to provide review and input.

The module builder, a visual editor for creating and modifying patient-generator modules using the GMF, was used as the authoring tool for module construction. States and transitions were iteratively modified and refined throughout the module-building process. The State Editor view enabled module developers to build the modules graphically without extensive knowledge of the underlying JSON representation; inside the module builder, the JSON source file could also be edited directly via the JSON Editor view. Edits made to the JSON source file were reflected instantly in the module's graphical diagram representation.

Once built, each newly developed module was downloaded as a JSON file and saved to the appropriate directory within the developer's local version of Synthea. The Synthea wiki[12] provides instructions for cloning a local build and configuring Synthea to generate data using a variety of run parameters as needed. JSON source files for new modules are loaded into a module developer's local version of Synthea. Using Synthea build processes, module developers generated synthetic health data and associated synthetic health records with specified run parameters and output in a variety of formats from the new modules.

Existing modules in the Synthea GitHub Repository[16] lack a consistent representation of supporting documentation for reviewers or implementers. So, as part of this project, a Module Companion Guide (companion guide) was compiled for each newly developed module. These companion guides now reside

in a centralized location within the Synthea wiki. Each companion guide provides additional detail about the Synthea module, including:

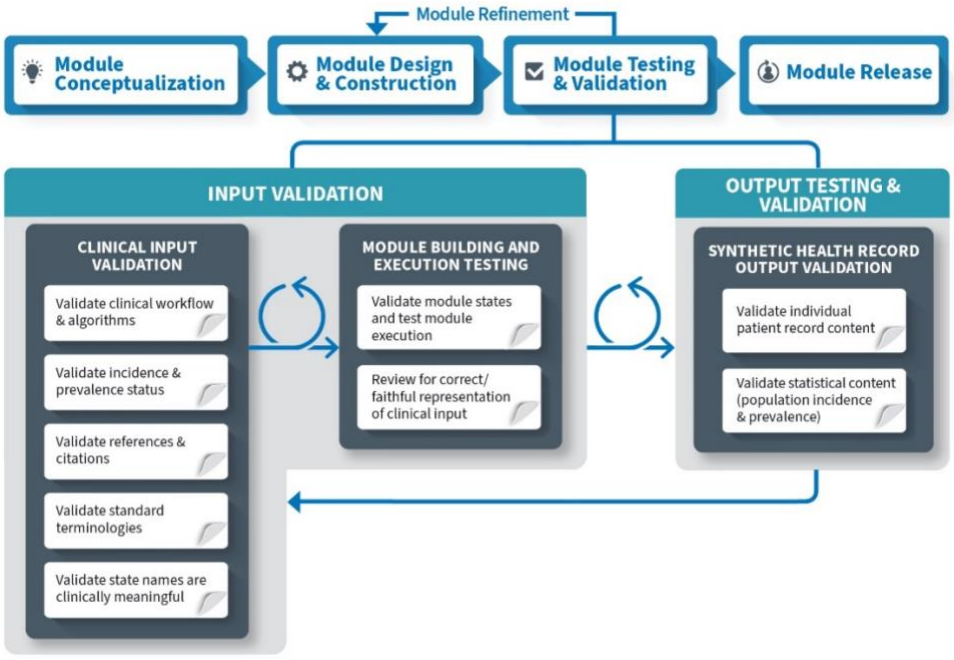
- A Module Parameters table, description, and universal disclaimer that provide a summary view of prevalence rates and explain the module's purpose.
- Sources and citations, such as clinical guidelines, scientific papers, health statistical data, and articles from reputable medical sites and government agencies, used as data sources to construct states and assign probabilities. For states where prevalence rates were not available, explanations for probability assignments were added.
- To assist with the validation of prevalence rates by the TEP and Synthea clinical reviewers, the Module States tables document how the probabilities are set for each state.

Module Testing and Validation

Each new module underwent iterative development and refinement that yielded increasing degrees of detail and accuracy.

Module testing and validation is an essential step in the module development lifecycle, occurring during and following module design and construction. Early in the development of the first modules, module developers found that testing and validation were inextricably intertwined with design and construction. Throughout testing and validation, each of the five modules (described in Newly Developed Synthea Modules) underwent iterative development and refinement that, over time, yielded increasing degrees of detail and accuracy. This iterative process was key to the overall quality and usability of the synthetic health data and associated synthetic health records generated by each module.

Figure 2: Module Testing and Validation



Clinical Input Validation

The primary goal of input validation is to ensure that the Synthea module under development is clinically valid and technically correct. Clinical reviews conducted by the TEP served as the modules' primary source of clinical input validation. To conduct effective clinical validations during TEP meetings and offline reviews, module developers created user-friendly module diagrams that supported review and discussion of clinical workflow and algorithms. Appendix A: Spina Bifida Visio Diagram provides one sample of the Visio module diagrams that were created through multiple iterations of module development. TEP meeting slides were designed to elicit questions and highlight key points related to the clinical flow and data sources requiring clinical input. Additionally, when a module pull request was submitted for inclusion of the new module in the Synthea GitHub repository, the module was evaluated by Synthea clinical reviewers using a standard Clinical Assessment Form, which provided a third-party clinical validation.

Module Building and Execution Testing

Once the module was built in the module builder, execution and testing were conducted to validate module states and confirm the accurate representation of clinical module input. Module developers viewed the module in the module builder to enable modification on an iterative basis. The Synthea developer setup allows developers to examine, extend, and/or build the Synthea source code locally. After a module is modified in the module builder, its source JSON file is downloaded and saved locally for execution and testing. Results from module testing and validation were fed back into module design and construction (Figure 2) to refine the module and improve its accuracy.

Execution testing involved careful inspection and resolution of warnings automatically generated by the module builder to ensure that the module used correct Synthea module states, transitions, and logic controls to represent the intended clinical care maps and statistics data. However, module builder warning capabilities are limited, and the Synthea wiki contains no information describing these warnings or how to resolve them. For example, code collision warnings appeared while building the Spina Bifida and Sepsis modules, alerting module developers that the same code was used by a state defined in an existing Synthea module. However, despite the warning, the module functioned as expected.

Careful visual inspection of states and transitions were conducted during module building to ensure proper state types and transitions were used.

Synthea initiates with a set of fictitious patients having predefined transition parameters and initial conditions, each evolving through discrete states over time according to probabilistic models or rules to produce an analytic population. Module developers relied primarily on visual inspection of the module in the module builder and the process of building and executing the module to test that proper state types (e.g., procedure, observation) and transitions were used. For example, the module developers found that warnings were not generated by the module builder when a clinical state is missing a code or when the low value of a "Delay state" range is higher than its set high value. These types of errors were discovered through careful visual inspection and execution testing.

Building and executing Synthea modules requires that they be free from runtime build errors. These errors were difficult for module developers to troubleshoot because their causes were not easily identified. When two common runtime errors, Java OutofBoundsException and Java NullPointerException, would occur,

module developers sought potential causes by visually examining the modules in the module builder. If visual examination did not reveal error sources, the generated synthetic health records were analyzed and the most recent JSON file was compared to the last error-free version of the module source file to help identify what may have caused the error. Module developers observed that certain runtime errors were associated with a specific state or only impacted a certain percentage of patients.

Synthetic Health Record Output Validation

Synthea-generated synthetic health records must accurately reflect the real clinical health data contained in the health records of actual patients. The properties of realism in synthetic health data generation support a greater degree of accuracy, reliability, effectiveness, credibility, and validity when data are used for PCOR.[18,19,20,21] Module developers' output validation testing for the new modules included validating synthetic health records and statistical content at the individual patient level. This ensured that each Synthea module used correct module states, transitions, and logic controls to represent the intended clinical care maps and statistics data defined within the module. For example, if a state was designed to be triggered only when a synthetic patient reaches a certain age, the data were checked to ensure that the specific state appeared only in the synthetic health records of patients of a certain age.

To ensure they accurately reflect real clinical health data, Synthea-generated synthetic health records must be validated for realism.

When running Synthea, module developers used specific configuration options to identify parameters that would generate appropriate output for different testing scenarios. This included which module(s) to run, number of synthetic health records to generate, and/or specific age ranges or gender to be represented in the data set. For example, the Prescribing Opioids for Chronic Pain and Treatment of OUD module is designed for patients 18 years of age and older, so specific run parameters were set to generate patients within an age range of 18 years and older and a population size of 200 (or other sizes, depending on testing requirements). When a transition had a low prevalence rate, a larger population size was required to test the module pathway. Once the module ran successfully in Synthea, the Synthea build process automatically placed all Synthea-generated synthetic health records into an output folder within the local version of Synthea based on default configurations set in the synthea.properties file. Synthetic health records are generated in a variety of formats, including text, HL7 FHIR®, comma-separated values (CSV), and HL7 C-CDA®, and are contained in a separate folder inside the output folder.

Module developers visually inspected the plain text output for accuracy as part of validating individual Synthea-generated synthetic health records. Some characteristics that were inspected included expected encounters, conditions, procedures, and/or properly generated medications. Timing relationships among data elements contained within the FHIR output were inspected to ensure that they were occurring as expected. For population-level validation, module developers used the CSV output and Excel pivot table functionality to analyze prevalence rates at the population level. Results from this validation are provided within each companion guide. Findings related to output testing and validation were shared and discussed with the TEP, and feedback was recorded and incorporated into the modules.

Module Testing and Validation Highlights

Module testing and validation may require many iterations and can be time consuming. Additionally, some conditions occur infrequently so, as a practical way to enable record generation and output evaluation, module developers temporarily increased the prevalence rate to a higher percentage. One complex issue discovered during the output validation of Synthea-generated synthetic health data involved symptoms not displaying in the generated FHIR or text output. Further testing revealed that symptoms will only display after a setting in the synthea.properties file is modified to generate a separate text symptom record for each synthetic health record. After consultation with the team at MITRE, the module was modified to use coded conditions for all secondary conditions, along with additional check simple states; this resulted in the conditions easily displaying in all forms of output.

Module testing and validation may require many iterations.

Modules were iteratively developed after reviewing their output. For example, when the number of deceased patients in the synthetic health records for the Spina Bifida module appeared high, module developers refined the module with more detailed mortality rates informed by an additional literature review. A review of sepsis patients' synthetic health records identified several clinical issues, most notably a lack of correlation between the systolic and diastolic blood pressures. Each systolic and diastolic blood pressure is randomly selected from a range set within Synthea; therefore, the pressures may not correspond to one another clinically. Additionally, the generated synthetic health record (Figure 3) alphabetically displays the diastolic value above the systolic value, rather than the systolic value above the diastolic value, as would be seen in a real patient's clinical health record. Similar issues were encountered while developing pain scales and corresponding medication treatments in the Prescribing Opioids for Chronic Pain and Treatment of OUD module.

Figure 3: Sample Sepsis Blood Pressure Reading in Text Format

```
OBSERVATIONS:
  2012-04-08 : Blood Pressure
             - Diastolic Blood Pressure                108.0 mm[Hg]
             - Systolic Blood Pressure                 45.0 mm[Hg]
  2012-04-08 : Lactate [Mass/volume] in Blood            3.1 mmol/L
  2012-04-04 : Blood Pressure
             - Diastolic Blood Pressure                79.0 mm[Hg]
             - Systolic Blood Pressure                 90.0 mm[Hg]
  2012-04-04 : Lactate [Mass/volume] in Blood            3.7 mmol/L
  2012-03-24 : Blood Pressure
             - Diastolic Blood Pressure                111.0 mm[Hg]
             - Systolic Blood Pressure                 107.0 mm[Hg]
  2012-03-24 : Lactate [Mass/volume] in Blood            2.2 mmol/L
```

NEWLY DEVELOPED SYNTHEA MODULES

The new Synthea modules, described below, targeted high priority use cases for patients with complex care needs, opioid use, and pediatric populations, and followed the standardized module development process described in previous sections. The modules were built by two experts specializing in data standards, health care interoperability, and clinical model design. Although neither had previous experience building modules using the Synthea GMF, they were able to utilize Synthea wiki documentation as a resource.[8]

Table 2: Newly Developed Synthea Modules

Module Title	Summary and Resources
Prescribing Opioids for Chronic Pain and Treatment of Opioid Use Disorder	• Models patients 18 years old or older. • Based on the Centers for Disease Control and Prevention (CDC) Guideline for Prescribing Opioids for Chronic Pain. [22] • The Treatment of Opioid Use Disorder, a fully developed component (submodule) of the main module (Prescribing Opioids for Chronic Pain), focused on an important issue associated with prescribing opioids and is based on the American Society of Addiction Medicine (ASAM) National Practice Guideline for the Use of Medications in the Treatment of Addiction Involving Opioid Use.[23] • Companion Guide and Module in Synthea
Treatment of Sialorrhea in Cerebral Palsy	• Models the treatment of sialorrhea in CP for patients 18 years of age or younger. • Based on the American Academy for Cerebral Palsy and Developmental Medicine (AACPDM) clinical care pathway for sialorrhea[14] and other publicly available data sources. • To provide realism in the synthetic health records, the module contains, but does not fully develop, other conditions that may occur in the CP patient, such as dystonia, pain, spasticity, central hypotonia, osteoporosis, epilepsy, gastroesophageal reflux disease, and intellectual disability. • Companion Guide and Module in Synthea
Sepsis	• Models the treatment of sepsis in patients 18 years of age or older. • Sepsis is a leading cause of death in critically ill patients in the United States.[24] • Based on the Surviving Sepsis Campaign clinical care guidelines for sepsis, including the guidelines for the Hour-1 Bundle for initial resuscitation in sepsis and septic shock.[25] • Companion Guide and Module in Synthea
Spina Bifida	• Models myelomeningocele (“open spina bifida”), the most severe form of spina bifida, for patients under 18 years old. • Spina bifida, a neural tube defect affecting the spine, is the most common, permanently disabling birth defect associated with life.[26] • Based on Guidelines for the Care of People with Spina Bifida[27] and CDC Spina Bifida.[28] • Companion Guide and Module in Synthea
Acute Myeloid Leukemia	• Models the parameters described in a microsimulation study by McCormick et al.,[29] which provides an evaluation of levofloxacin use in children with AML. • To test the utility of Synthea synthetic health data for simulation studies, the module was developed to replicate the original study’s published parameters; this created the module’s initial conditions and simulation endpoints of demographics, health events, costs, and mortality. • Companion Guide and Module in Synthea

Challenge Competition

The purpose of the Synthetic Health Data Challenge (Challenge) was to increase awareness and understanding of Synthea and its capabilities and to engage a broader community of researchers and developers on which open-source programs like Synthea rely. Innovators, researchers, and developers from across the nation contributed to the creation and testing of solutions aimed at further cultivating the capabilities of Synthea and the synthetic health data it generates.

The Challenge was formulated to inspire a wide range of experts to demonstrate novel uses of Synthea and validate the realism of Synthea-generated synthetic health data.

The Challenge occurred from January 19, 2021 to July 13, 2021 and was conducted in two phases. For Phase I – Proposal for Innovative Models, participants described their proposed solutions, including methodology and intended outcomes. Selected proposals moved on to Phase II – Prototype/Solution Development. Winning solutions were selected from among the Phase II submissions. Solutions had to be in one of two categories (Category I – Enhancements to Synthea or Category II – Novel Uses of Synthea Generated Synthetic Data) and address at least one of the three PCOR priority use case focus areas. A validation component was also required.

Technical reviewers with expertise relevant to the Challenge reviewed all Phase I and Phase II submissions to ensure they met Challenge criteria. Each submission was then evaluated by federal employees serving as Challenge judges and an ONC employee serving as an alternate judge. Judges had knowledge and understanding of synthetic health data, PCOR research, and health IT standards and tools.

- For Phase I Proposals, judges awarded points based on 10 criteria crafted to evaluate each submission's potential for further development.
- Phase II Solution Packages were scored on six aspects of Impact and Innovation, six aspects of Functionality and Implementation, and three aspects of Validation. Optional bonus points were awarded for outstanding contributions to Synthea.
- Judging was subject to a final decision by the Award Approving Official: Micky Tripathi, PhD, National Coordinator for Health Information Technology.

Twelve Phase I Proposals for Innovative Models were received. Phase I Challenge entries were submitted by experts from across the United States with diverse professional backgrounds, including analytics, engineering, pharmaceuticals, digital innovations, academia, medicine, health informatics, data science, and applications development. One proposal was disqualified, and eleven were evaluated by the judges. Nine proposals were selected to move on to Phase II. Those participants were asked to build and test their solution and then present results in a five-page paper and a five-minute presentation.

The Challenge awarded a total of $100,000 in prizes, and winners presented their solutions during the Winning Solutions Webinar on October 19, 2021. Phase II Challenge winners were from many organizations: Allina Health, Battelle Memorial Institute, Komodo Health, LMI Consulting LLC, Mercy

Hospital St. Louis, Microsoft Corporation, Particle Health, UnityPoint Health, University of California San Francisco, University of Illinois Hospital & Health Sciences System, Utica College, and Vanderbilt University Medical Center. Winning solutions are described in Table 3.

Table 3: Phase II Winning Solutions

	TEAM & SOLUTION TITLE	DESCRIPTION
1st Place Winner $40,000	CodeRx: *Medication Diversification Tool*	The medication diversification tool (MDT) enhances the clinical realism of Synthea-generated synthetic health data by increasing the variation of medication orders in the Synthea GMF. MDT increases the diversity of medications prescribed based on three publicly available government data sets, creating more practical synthetic medication profiles in Synthea modules. The MDT can generate medication orders for all current or future disease modules within Synthea that are treated with a medication.
2nd Place Winners $15,000	Generalistas: *Virtual Generalist – Modeling Co-Morbidities in SyntheaTM*	To optimize Synthea functionality and application to the real world, the Virtual Generalist generates a wide range of complex combinations of conditions (co-morbidities) in statistically appropriate distributions. The Virtual Generalist operates by modifying attributes used by other modules, acting as a statistical supervisor to adjust population distributions.
	LMI: *On Improving Realism of Disease Modules in SyntheaTM – Social Determinant-Based Enhancements to Conditional Transitional Logic*	Greater realism in the Prescribing Opioids for Chronic Pain and Treatment of OUD module was produced by improving characterization of features related to key social determinants of health (SDOH) and creating new software functionality that enables users' access to these features when developing conditional transition logic in disease modules.
3rd Place Winners $10,000	Particle Health: *The Necessity of Realistic Synthetic Health Data Development Environments*	The Particle Health Sandbox addresses the absence of practical synthetic health data testing environments with C-CDA documents and in-document provider notes. The Sandbox environment includes modified Synthea-generated synthetic health records in the form of C-CDA documents with accompanying point-in-time documents, in-document chart notes, FHIR® documents, and pre-loaded patient types—including opioid and COVID-19 synthetic health records generated by Synthea modules.

	TEAM & SOLUTION TITLE	DESCRIPTION
	Team TeMa: *Empirical Inference of Underlying Condition Probabilities Using Synthea™-Generated Synthetic Health Data*	To determine the most likely patient pathologies based on a given set of observed symptoms and patient demographics, two approaches were implemented. Team TeMa first evaluated synthetic health records generated by Synthea to study the complicated relationships between sets of symptoms and the likelihoods of possible underlying causes. The second approach used synthetic health data to populate probability distribution functions in a graph-based machine learning model and empirically analyze them to obtain posterior pathology probabilities.
	UI Health: *Spatiotemporal Big Data Analysis of Opioid Epidemic in Illinois*	The spatiotemporal distribution of emergency 911 calls and ambulance dispatches related to opiates overdoses in Chicago were explored using the Prescribing Opioids for Chronic Pain and Treatment of OUD module. Synthea-generated latitude/longitude (x,y coordinate) data were matched with Esri™ demographic data and enhanced with SDOH information to generate public health-related prediction models for clinical outcomes.

Challenge competitions are a tool ONC has used to call on a wide array of stakeholders nationwide to help solve complex problems. ONC publicized this Challenge through targeted dissemination of press releases and messaging via LinkedIn and Twitter. The Challenge.gov Facebook page and Twitter feed also informed potential Challenge participants and the public about the Challenge's purpose and activities. This resulted in information about the Challenge and Synthea appearing nationwide in health IT news sources—including Health IT Analytics, Healthcare IT News, and FedHealthIT.com—as well as news sources for health care related public, private, and governmental organizations. The Challenge was also referenced in articles about synthetic health data in The National Law Review and The Wall Street Journal. Additionally, conference presentations disseminated information about PCOR, Synthea, and the Challenge, as well as generating follow up inquiries by industry stakeholders, including researchers. Presentations at the 2020 AMIA Annual Symposium, the 2020 ONC Annual Meeting, and the 2021 ONC Annual Meeting shared information about the Challenge and introduced the new modules in the Synthea tool to the community. In addition, a detailed demonstration of one of the newly developed modules (CP) was presented at the 2021 ONC Annual Meeting Health IT Infrastructure Tech Showcase.

In the first weeks of the Challenge, the Phase I Informational Webinar was attended by 116 people representing federal agencies, technology vendors, health care organizations, research facilities, and educational and nonprofit organizations. It provided information about participating in the Challenge; explained the purpose, benefits, and drawbacks of clinical and synthetic health data; and described Synthea

technology and requirements. A Phase II Informational Webinar was conducted exclusively for Phase II participants to provide detail regarding the requirements of participating in Phase II.

Approximately 150 inquiries were received in the dedicated Challenge email inbox. Responses to those inquiries were shared with participants in each phase of the Challenge via the FAQs and Technical Guidance and Tips on the Challenge.gov webpage.

Phase I Challenge entries engaged numerous individual experts with a wide range of diverse backgrounds.

The Challenge culminated with a Winning Solutions Webinar, featuring presentations by the six Challenge winners about their innovative work. The Winning Solutions Webinar was attended by 110 participants representing 78 organizations from six different countries.

Demonstration Study

The Synthea platform, originally designed for generating realistic synthetic health data for software testing, implements publicly contributed clinical modules representing life cycle and disease and treatment progression. This demonstration study (demonstration) evaluated whether Synthea is also suitable for simulation studies by assessing whether Synthea could be used to replicate a published simulation without modifying source code. The demonstration was an investigation that evaluated the scope and utility of Synthea for PCOR hypothesis testing. For the demonstration, the project team:

- Conducted an independent and comprehensive review of Synthea-generated synthetic health data, features, and gaps;
- Evaluated the utility of Synthea-generated synthetic health data for PCOR studies;
- Cataloged opportunities to improve Synthea functionality to support simulation-based hypothesis testing; and
- Cataloged opportunities to improve the design of Synthea modules to support future studies.

Simulations are used in health care research when desired empirical data are not available, particularly when uncertainty and unexplained variation might impact outcomes. The demonstration evaluated Synthea as a platform for simulation studies of patient health care utilization and outcomes. Such studies produce a simulated data set for statistical analysis comparing alternative treatment or policy scenarios or forecasting alternative future circumstances, including outcomes such as disease progression and mortality. The demonstration evaluated whether Synthea is suitable for simulation studies by creating a Synthea module that replicates a recently published (2020) simulation. It also assessed whether features of Synthea were sufficient for replicating the simulation without modification to the source code.

Synthea software generate synthetic health data in a format compatible with international standards for electronic health records (EHRs) and documents and enables users to create synthetic health data with ecological validity by supporting model-driven data generation. Synthea can generate a large volume of data in a format that can be immediately incorporated into sandbox environments for health care application testing. Secondary potential use cases include Synthea's use of Census data to generate a distribution of initial demographic-specific conditions reflecting local populations (a typical feature of microsimulation studies); a user interface that guides users through drag-and-drop model building, encouraging health care-based data elements and state transition modeling (best suited for simple models); and generation of realistic trajectories of data values over time (provided adequate investment in model development).

The project evaluated whether Synthea is suitable for simulation studies, often used by health care researchers.

Most Synthea modules are developed based on a combination of practice guidelines, expert input, and publicly available information. This method of module development is useful for software testing, but presents limitations for PCOR hypothesis testing. Using established clinical guidelines to create models does not provide data depictions of clinical practices outside of the guidelines (e.g., untested or comparatively less effective or agreed upon treatment pathways). Synthea was not designed to ingest real

clinical health data for parameter estimation. Rather, its parameters are derived from publicly available data and/or expert input.

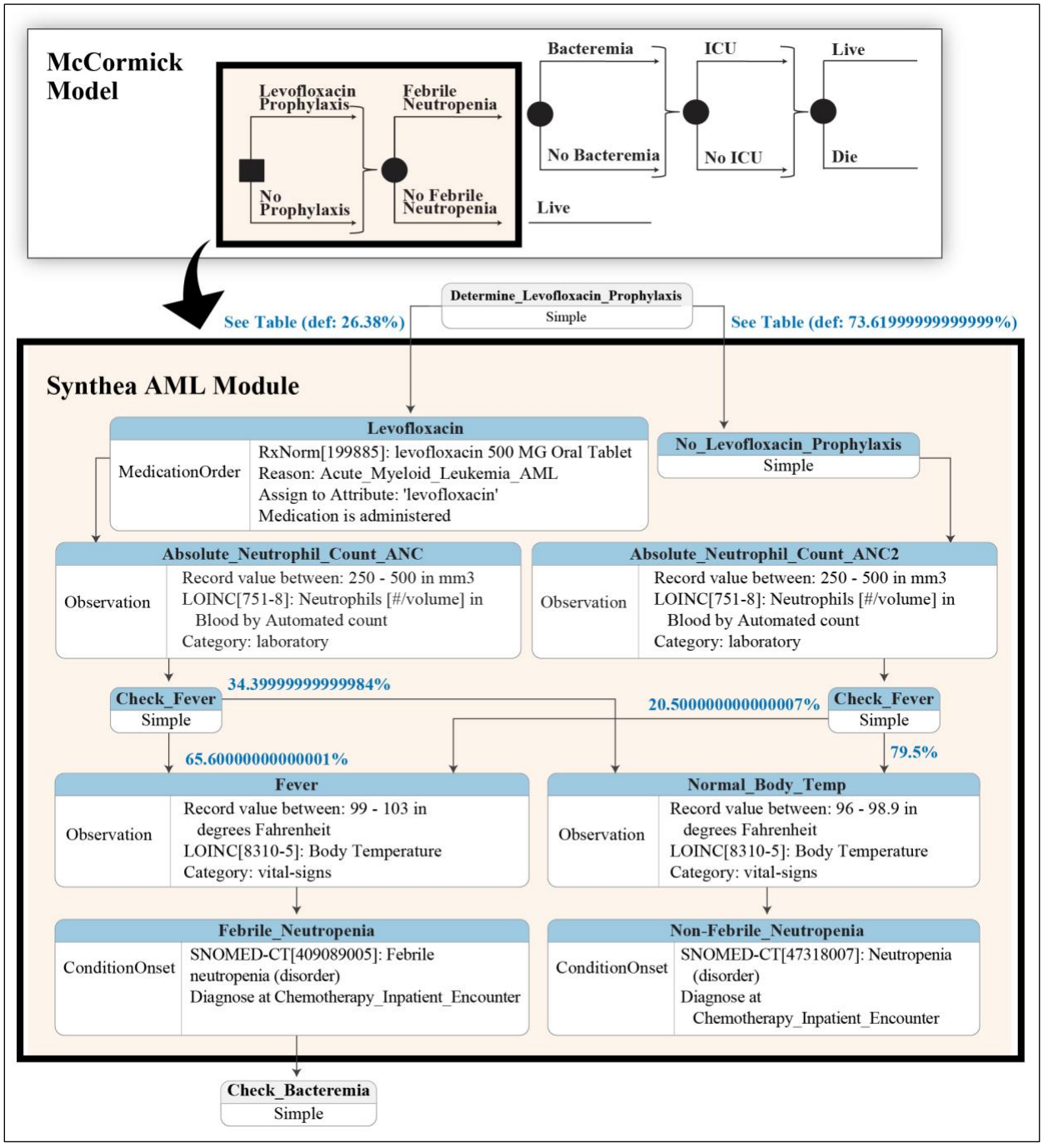
METHODS AND APPROACH

The project team conducted a literature scan to identify simulation studies with published endpoints and parameters for initial conditions, health states, transition probabilities, and service utilization. After evaluating published studies with models that could support module design and demonstration study analysis, a study by McCormick et al. [29] (McCormick study) was selected. The clinically relevant McCormick study provides an evaluation of levofloxacin use in children with AML. The McCormick study model is simple and straightforward with decision branches that readily align with Synthea design parameters. The model occurs within one clinical episode, which facilitates Synthea modeling because Synthea requires that all clinical treatments and procedures be modeled within one or more clinical encounters. The McCormick study decision-analysis model was designed to evaluate the effectiveness of levofloxacin prophylaxis compared to no prophylaxis during a single chemotherapy cycle in patients less than 21 years of age with AML. The study reports outcomes, including the cost of bacterial infection, cost per ICU admission, and cost per death avoided.

Evaluating the utility of Synthea-generated synthetic health data for PCOR hypothesis testing required a Synthea module designed with a microsimulation-based hypothesis in mind. The AML module was selected for development based on the use case evaluation criteria (Table 1) and because the McCormick study model aligned with Synthea parameters, as described above. The module was developed based on parameters from the McCormick study, which defined transitions between the module states. Components of the McCormick study initial population were used to help define select characteristics of the initial patient population within the Synthea module (Figure 4). Along with the parameters and population-level data, additional McCormick study assumptions were incorporated into the module design. For example, the McCormick study assumed bacteremia was only present in the setting of febrile neutropenia and routed non-febrile neutropenia patients to a terminal state.

The AML module was initialized with 25 states and 24 transitions. During module development, 22 additional states were incorporated, primarily for age delay, to replicate McCormick study age distributions. McCormick study pathways and assumptions were incorporated into the AML module. Transitions between states and distribution of attributes were based on McCormick study parameters. This approach varied from typical Synthea module development, which relies on clinical care guidelines and publicly available incidence and prevalence statistics. Workarounds were developed to address limitations of the Synthea modeling interface, resulting in an increase in the number of states and transitions. The AML module was finalized with 47 states and 46 transitions.

Figure 4: Comparison of McCormick Model to Synthea AML Module



To reduce the number of states and transitions, the module developers attempted to add a Gaussian distribution to one age delay at the top of the module. This functionality is new in Synthea, and it led to a certain percentage of patients being diagnosed and entering the encounter prior to birth (sometimes years before birth) to replicate the normal curve for age. Although this distribution with negative results would be acceptable for certain components of the clinical record, such as lab results, a patient is not able to receive

care prior to their date of birth. The issue was reported to the MITRE team. An initial fix was applied that removed the negative age patients; however, it led to an incorrect mean for the distribution. The project team continued working with the MITRE team to ensure the distribution will work appropriately in the future. Due to publication time constraints, the functionality was not included in the final published version of the AML module.

RESULTS

After the AML module was enhanced based on initial output validation, population-level validation was conducted. Module developers generated representative samples of Synthea-generated synthetic health records to validate incidence and prevalence rates compared to the McCormick study population-level data. This process used statistical analysis software (e.g., Stata or SAS) for a comprehensive and efficient analysis of the data, including statistical tests comparing the distribution of Synthea states to published results. Analysis methods and results were documented and, based on feasibility, additional module enhancements were applied. Representing custom demographic distributions that differed from Synthea's default demographics (which are based on Census data, costs, and non-uniform distributions of continuous variables), required developing and testing six module versions and multiple iterations of each version.

The final stage of analysis replicated the McCormick study simulation results and hypothesis tests using the AML module. Point estimates in McCormick study tables and supplemental material were compared to Synthea data using chi-squared and t-tests for binary and continuous variables, respectively. Pass-fail tests for each iteration were assessed and updates made until all comparisons "passed" with $p>0.001$. Issues were tracked, and module builders identified and implemented solutions.

Table 4 compares results between Synthea and the McCormick study for each sub-iteration of AML module V0.6. Matching the Gaussian distribution for age was the most challenging aspect of module development and required multiple attempts to replicate reference values over the course of development.

Table 4: Summary of Results for AML Module V0.6

AML Module Iteration	V0.6a		V0.6b		V0.6c		V0.6d		V0.6e (w/ age restriction)		V0.6e (w/delays)	
Treatment (Tx) vs. Standard of Care (SoC)	**Tx**	**SoC**	**Tx**	**SoC**	**Tx**	**SoC**	**Tx**	**SoC**	**Tx**	**SoC**	**Tx**	**SoC**
Population	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
Age (SD)	fail	fail	fail	fail	fail	fail	fail	fail	fail	fail	pass	pass
Race: White	fail	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
Race: Black	fail	fail	pass	pass	fail	pass	pass	pass	pass	pass	pass	pass
Race: Other/Missing	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
Hispanic Ethnicity	fail	fail	fail	fail	fail	pass	pass	pass	pass	pass	pass	pass
Bacteremia	fail	pass	fail	pass	fail	pass	fail	pass	fail	pass	pass	pass
ICU	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
Mortality	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass

The level of effort to develop the AML module was approximately 20 percent higher than other modules developed by the same team. The AML module required a unique design approach and the expertise of a research analyst who examined replication data. Accommodating McCormick study parameters and a Gaussian distribution required creating numerous delay states and complex table transitions. Achieving a statistically precise level of accuracy when comparing the Synthea output to McCormick study results required developing and testing numerous iterations of the AML module.

The demonstration study revealed that Synthea can be repurposed for simulation studies comparable to the McCormick study without advanced programming. Although some workarounds were required, these could be accommodated with module-specific lookup tables or introducing an array of states and uniform transition probabilities to approximate normal distributions.

The McCormick study simulation did not involve multivariate models for conditional state transitions or interactions between individuals. However, it is representative of a broad class of microsimulation studies. Future releases of Synthea may include user-friendly support for these features, extending the complexity of possible studies and models.

Synthea can be used to support study planning, software testing, and development of analysis routines for research studies.

Demonstration study results also suggest that Synthea can be used for study planning, software testing, and development of analysis routines for research studies. For complex experimental designs, such as stratified or cluster randomized trials, simulation study data are often used in power analysis to prepare for clinical trials. Synthea can generate data for power analysis and sample size requirements. For example, should a follow-up to the AML simulation include a randomized trial, investigators can apply Synthea to help plan sample size requirements for different magnitudes of levofloxacin effectiveness. Synthea can generate realistic

health care data, comparable to data generated in EHRs, during pragmatic trials. Informaticians preparing for research studies can generate and use data to apply and test software that may, in a forthcoming trial or registry project, electronically extract data from EHRs. Once a database is designed with realistic data, analysis routines for data safety and monitoring reports can be developed and tested before beginning data collection, reducing the time between implementing research infrastructure and initiating enrollment.

Informaticians preparing for research studies can use Synthea-generated data to test software that may electronically extract data from EHRs.

For health economics research, Synthea can represent costs as they may appear in billing data for encounters, procedures, medications, and immunizations using the billing codes provided in Synthea's lookup tables. The McCormick study assigned a generic cost to each outcome and medication; thus, cost comparisons can be calculated with these frequencies. Future analyses and module iteration could add this granularity and complexity and assess or extend Synthea's default values for costs. Although Synthea allows for more extensive modeling, the current strategy is sufficient for study designs like the McCormick study, where simple frequencies of generic simulated events are used to generate cost comparisons.

Findings and Lessons Learned

MODULE DEVELOPMENT, TESTING, AND VALIDATION

- The TEP provided invaluable "real world" clinical feedback and problem-solving that enhanced overall module quality. These experts fully supported the project goals of providing high quality, realistic synthetic health data for researchers, health IT developers, and informaticians.
- Module Companion Guides serve as a useful tool for future developers or users of a Synthea module. They provide detailed documentation, including metadata, module diagram, module states, module parameters, and sample population-level Synthea output results.
- Standardizing the Synthea module development process supported the selection of candidate use cases and the development and enhancement new modules. The module development methodology outlined a consistent strategy for all modules, beginning with conceptualization, moving iteratively through design and construction, testing and validation, and culminating with the new modules' inclusion in the Synthea GitHub repository for use by the community.
- The iterative module development process yielded increasing degrees of detail and accuracy and was key to the overall quality and usability of the synthetic health data that the new modules produce. The process of module testing and validation is inextricably intertwined with module design and construction.
- The graphical view of each new module was large, complex, and difficult to read. Simplified module diagrams supported a more user-friendly view for clinical reviewers to validate clinical workflow and algorithms as the modules were developed and refined.
- Static images of modules can be viewed in the Module Gallery.[30] Users can also view, interact with, and edit the most current version of all modules using the module builder.**Error! Bookmark not defined.**

ENHANCING COMMUNITY AWARENESS OF SYNTHEA

- Publicizing the Challenge at launch was essential to engaging a pool of diverse, professional participants from across the United States. Continued dissemination efforts kept the attention of non-participating stakeholders, resulting in news stories of national interest.
- The Winning Solutions Webinar rewarded the efforts of the Challenge winners by allowing them to present their innovations to a national audience. By highlighting their activities, the winners spurred interest in their specific endeavors, added to the ongoing conversation regarding the role of synthetic health data in health research and health IT development, and, perhaps more importantly, served to strengthen interest in Synthea as a useful, adaptable tool for PCOR research.

- The Challenge led to unexpected results from unlikely stakeholders while generating enthusiasm and cultivating capabilities within the community. Proposals offered novel approaches and innovative solutions with ideas ranging from medication diversification tools to incorporating SDOH to realistic synthetic health data sandbox environments.
- Public presentations about the project during the 2020 AMIA Annual Symposium and 2021 ONC Annual Meeting generated additional interest in Synthea and its capabilities. Follow-up meetings were held with individual organizations to discuss potential uses of Synthea and its synthetic health data in academic research centers.

DEMONSTRATION STUDY

- At the time of the demonstration study, Synthea lacked the capability to support complex multivariate distributions. However, newer functionality supports non-uniform distributions, including univariate Gaussian distribution and stochastic transitions; this may offer greater flexibility and support for module development. Enhancements to documentation of Synthea's existing features may facilitate future simulation research without significant changes to source code.
- For health economics research, Synthea can represent costs as they may appear in billing data for encounters, procedures, medications, and immunizations using the billing codes provided in Synthea's lookup tables. The McCormick study[29] assigned a generic cost to each outcome and medication, thus showing that cost comparisons can be calculated with specific frequencies. Future analyses and module iterations might add to this granularity and complexity and assess or extend Synthea's default values for costs.

Considerations for Advancing Synthea for PCOR

The work conducted by this project confirmed that, with community engagement, Synthea can be iteratively improved over time to enhance content, generate more accurate and realistic synthetic health data, and expand research applications. The improvements implemented by the project team were supported by counsel and input from researchers, clinicians, project partners, and other stakeholders, thereby laying a foundation for ongoing collaboration between the synthetic health data community and researchers. Since its inception, Synthea's evolution as an open-source platform has depended on community involvement. Continuing to cultivate the perspectives and stewardship of this project's actively engaged users and contributors will help ensure Synthea's continuing robust growth, evolving functionality, and capacity to serve all PCOR stakeholders. Based on the project team's experiences, this section describes opportunities for future work that will enhance Synthea software and guidance and support PCOR researchers.

OPPORTUNITIES TO ENHANCE SYNTHEA SOFTWARE

Enable Synthea modules to more realistically depict the real world by expressing variations in care and data quality that occur in real health records.

Synthea modules are often built using clinical care guidelines and standards of care. As a result, Synthea-generated synthetic health records lack the variations in care and data quality that regularly occur in real health records. Without this level of clinical reality, the data sets produced by Synthea limit researchers' ability to test algorithms and tools that seek to identify variations in care. The open-source community should explore opportunities to further enhance the Synthea GMF to introduce variations in care and data quality into Synthea-generated health records.

Expand Synthea to focus beyond care provided in clinical settings.

Synthea currently supports the ability to represent treatments occurring in a clinical setting. The ability to represent treatments occurring outside a clinical setting or address behavioral and neurodevelopment disorders would support modeling conditions such as autism and, over time, contribute to more patient-centered synthetic health records. For example, treatments such as Applied Behavior Analysis, Discrete Trial Training, alternative medicines, and dietary approaches, which take place in the community and home, would support expansion of Synthea modules and the synthetic health data Synthea generates.

Add SDOH to Synthea modules.

The ability to add SDOH factors to Synthea modules would be beneficial; adding branching logic would enable a more representative simulated population. The relationship between SDOH factors and health disparities, inequities, and resulting health outcomes could be supported by this functionality. For example, the addition of data types such as education access and quality, economic stability, and health care access would enhance Synthea synthetic health data to support SDOH use cases. As USCDI evolves, Synthea

output should also evolve to incorporate new data elements that will provide researchers with more granular data. For example, newer iterations of USCDI include data elements like sexual orientation, gender identity, and SDOH to help address disparities in health outcomes for minoritized, marginalized, and underrepresented individuals and communities.[31]

Enable Synthea to perform mathematical calculations.

The ability to perform mathematical calculations based on randomly generated data would be a valuable enhancement to Synthea. For example, Synthea currently randomly generates the mean scores of the Pain, Enjoyment of Life, and General Activity (PEG) scale within a prescribed range, rather than calculating a mean using Synthea-generated individual item scores. This issue also impacts blood pressure readings, a problem that the project team encountered while developing the Sepsis module. Systolic and diastolic pressures are generated randomly from a range, so pressure differences may not correlate or express a realistic blood pressure reading.

Make replicating population-level statistics in Synthea easier.

Each time Synthea is run, the random nature of the simulation produces different output. This poses a challenge when users are attempting to replicate the same output for testing purposes. Although Synthea offers an -s run parameter configuration to generate a population using the same seed, the project team discovered that, even when using the same version of Synthea, the generated synthetic health records were not always identical, particularly when generating a large sample. It would be useful if Synthea could allow users the option of configuring the software to run the same simulation repeatedly and in large numbers.

Expand Synthea's use of specific terminology and code systems.

Synthea uses specific terminology and code systems (e.g., SNOMED-CT, LOINC), but does not currently support other code systems, such as ICD-10. This prohibits certain module use cases from being contributed and limits the realism of the synthetic health records. Consideration should be given to expanding Synthea to accept code systems commonly used in EHRs.

Enable Synthea to assign physician specialties.

Within Synthea, every clinician is designated as a general practitioner. Expanding the provider specialties within Synthea to allow simulation of patients who visit specialty practitioners would enhance the reality of Synthea-generated synthetic health records. For example, a patient with a cardiac condition might see a cardiologist on a routine basis.

Make module remarks viewable before modules are opened.

Module remarks function as metadata to a module. Currently, module remarks can only be viewed, one module at a time, using the module builder in the module remarks section or by opening the source JSON file directly. Making module remarks more accessible and allowing users to simultaneously view the module remarks for all modules would serve as a catalog to help new users more easily find modules of interest. The current module remarks field is an optional free text field. Consideration should be given to making the module remarks section structured and required to provide better metadata, resulting in improved consistency across modules.

Provide a database schema for Synthea.

Synthea generates synthetic health records in CSV format, such as conditions.csv, encounters.csv, and patients.csv. These CSV files are important for population-level validation and data analysis. The CSV files can be imported into an Excel spreadsheet to create pivot tables or imported to a database for advanced analytics or other usage; data in each CSV file then acts as a table in a database. However, the Synthea wiki does not provide a database schema that shows the relationships of these tables. Providing a database schema of Synthea tables (e.g., conditions, encounters, patients, medications), will help users visualize how these Synthea tables are constructed, what fields are contained in a table, and what fields serve as primary or foreign keys.

Enable Synthea to produce digital X-ray images.

Synthea currently generates synthetic health records indicating that imaging procedures were performed. Results are generated using the ImagingStudy state and appropriate DICOM codes, however, Synthea is unable to simulate digital images. Synthea-generated synthetic health records would be improved and enhanced by adding the ability to produce Digital Imaging and Communications in Medicine (DICOM) format images to the records.

OPPORTUNITIES TO ENHANCE SYNTHEA GUIDANCE

Enable Synthea to run a single module for testing and development purposes.

Running a single module for testing and development purposes is difficult. Even when using the -m function, extraneous information from default modules is included in the output. (Note: In the time since the new modules were developed, information regarding the -m function has been removed from the Synthea wiki page.) This functionality is important to the Synthea software tool because it allows module developers to test the new module and ensure it runs appropriately. Additionally, many users of Synthea request the ability to generate synthetic health records with only one or two conditions. Given the importance of the -m function to new module developers and other users, consideration should be given to adding this back to the Synthea wiki along with clarifying that some default module content might still be included in the output.

Improve documentation for handling errors in the Synthea module builder.

The Synthea wiki has limited information for handling errors in the module builder and when generating synthetic health records. This can make testing difficult and often requires a “try it and see” approach. Some errors are warnings, such as code collision warnings, while other errors cause failures when generating synthetic health records. Consider adding more information to the Synthea wiki to explain warnings and errors within the module builder and when running Synthea to generate synthetic health records. This information could be helpful to novice users testing newly developed Synthea modules.

Provide documentation that promotes standardization of naming conventions for module states and codes as well as uniform placement of references and citations.

Currently, there is no standard naming convention for states or suggested placement for references and citations within Synthea modules. Additionally, there is no standard for naming codes within states, such as when to include the SNOMED descriptor, which varies across modules and can impact the generated synthetic health record output. Consideration should be given to adding documentation to the Synthea wiki recommending a standard naming convention for all module states and module state codes. Additionally,

adding documentation suggesting placement of citations within modules could support module developers building on existing modules.

Provide enhanced guidance for displaying information in varied output formats and generating symptoms and encounter times in the synthetic health data output.

Certain items, such as timing for encounters, only show up in FHIR® output. Other items, such as symptoms, do not display in the current text format or FHIR format and will only generate in a separate text file. Information clarifying this should be added to the Synthea wiki. In addition, consideration should be given to adding information to the Synthea wiki to guide users to use coded conditions as often as possible if they want a condition to display in all forms of output. The Synthea wiki should also instruct users to generate symptoms and encounter times in the output. This information will assist novice and experienced module builders as they develop and test new modules.

Provide module testing documentation and strategies in the Synthea wiki.

Certain strategies must be employed during testing to ensure that conditions and related treatments generate in the output for testing purposes. One such strategy, used by the project team, was temporarily increasing the prevalence rate to a higher percentage for testing low probability conditions. Adding information to the Synthea wiki on testing newly developed modules along with helpful testing strategies would benefit module developers as they test new modules.

Enhance Synthea wiki documentation with information about the new table transition in the Synthea module builder.

The new table transition in the Synthea module builder must be redeveloped each time the module is uploaded into the module builder. Additionally, the table must be created very specifically with rows totaling 1.0, without white spaces, and with commas in between column headers. Enhanced documentation in the Synthea wiki regarding this new function would support module developers. For example, added information might explain that the table must be recreated each time a module is uploaded into the module builder for editing. Also, once a CSV file has been created and downloaded from the Synthea module builder, the table CSV file must be uploaded separately into the Synthea module's lookup table folder in a developer's local version of Synthea for the module to run appropriately.

OPPORTUNITIES TO SUPPORT PCOR RESEARCHERS

Provide additional resources for users who are not proficient with Synthea technology and tools.

Adding a dedicated location that enables easier access to documentation and correspondence with Synthea developers could support users who are interested in using Synthea and Synthea-generated synthetic health data, but unfamiliar with Synthea technology and tools. For example, although the Synthea module builder user interface is user friendly, some users may need additional support with navigating GitHub or resolving JSON warning messages when using Synthea to generate synthetic health records. The Synthea wiki is comprehensive, but some gaps in documentation exist. Creating a formal website for Synthea documentation (outside of the wiki) would help users easily access and view Synthea documentation and communicate with Synthea developers.

Enhance documentation of Synthea's non-uniform distributions to facilitate future simulation research.

Newer Synthea functionality supports non-uniform distributions, including Gaussian distribution and stochastic transitions. This capability promises greater flexibility and support during module development and may facilitate the use of Synthea for simulation research; however, additional documentation about the availability and use of non-uniform distributions should be added to the Synthea wiki. This information will assist module developers and researchers who wish to develop and test these functions.

Support additional analysis of cost comparisons to further assess and extend Synthea's default values for costs.

Users can represent costs for encounters, procedures, medications, and immunizations by using the billing codes provided in Synthea's lookup tables. Future analysis of cost comparisons based on various module configurations and frequencies could enhance the granularity and complexity of Synthea output and extend its default values for costs. This improvement could be particularly useful for health economics research.

Conclusion

Synthetic health data can reflect the characteristics of a population of interest and be a useful resource for researchers, health IT developers, and informaticians. Researchers and developers often depend on clinical data, but high-quality real clinical health data can be difficult to access because of a variety of restrictions. Additionally, interoperability issues impede gathering data from different resources for robustly testing analysis models, algorithms, or developing software applications. Synthetic health data helps address these issues and speeds the initiation, refinement, and testing of innovative health and research approaches.

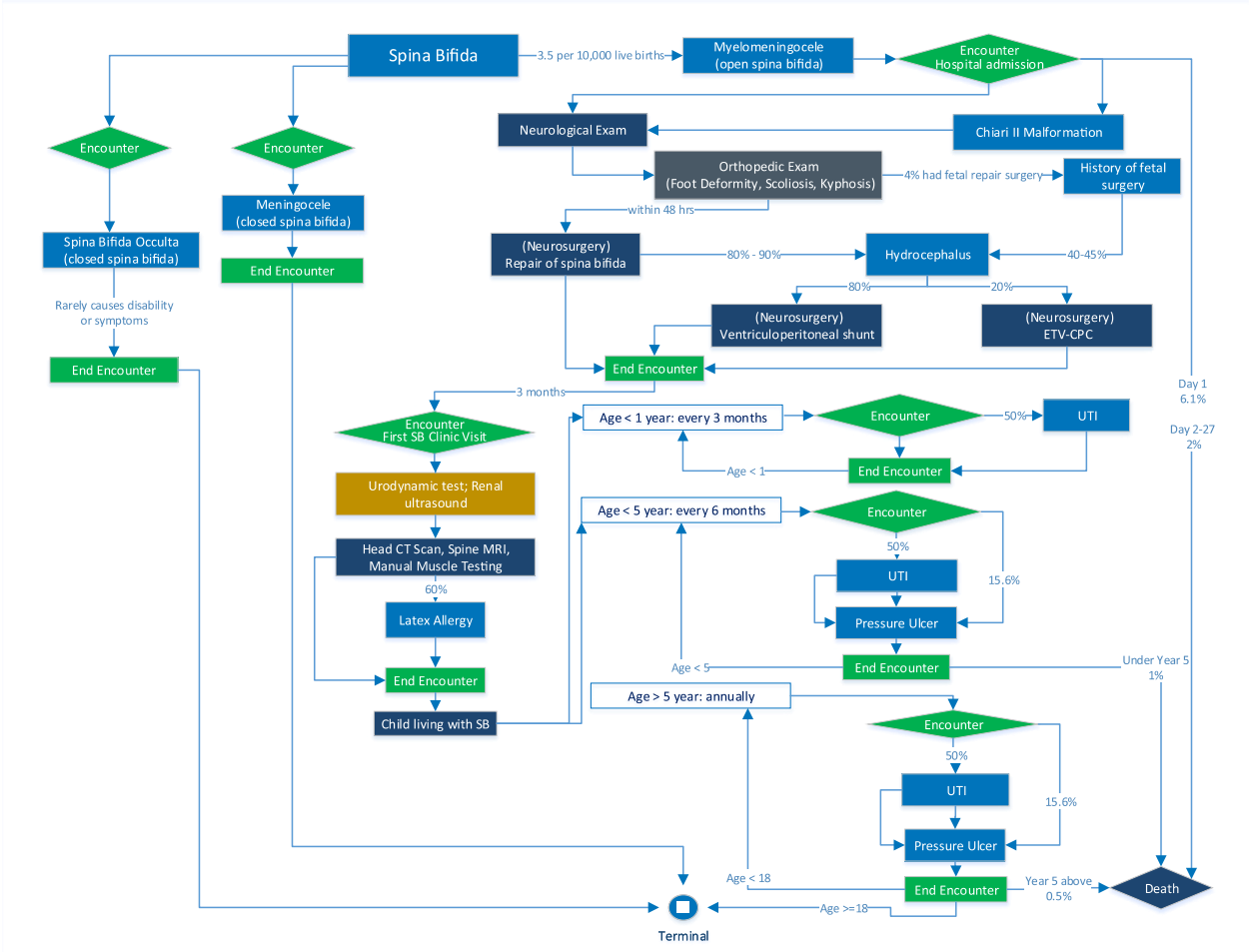
This report details ONC work that supports the generation of synthetic health data for research and health IT development. ONC led this effort to enhance Synthea, an open-source synthetic health data generation engine, to accelerate PCOR and other research. Increasing the availability of synthetic health data allows researchers to complement their use of real clinical health data and enhance their ability to conduct rigorous analyses, benchmark algorithms, and validate early hypotheses testing. Facilitating the use of synthetic health data for hypothesis and technology testing also supports the HHS objective of protecting the privacy of personally identifiable information. The availability of reliable and robust synthetic health data generation engines safeguard patient privacy by supporting appropriate stewardship practices in which real patient data are only accessed and used when necessary.

Appendices

APPENDIX A: SPINA BIFIDA VISIO DIAGRAM

A Synthea™ module diagram within the Synthea Module Builder is often large and complex to view, as it includes both clinical states and control states. It may be challenging for users to understand and navigate the module within Synthea, especially those who are new to the process. The purpose of providing Visio diagrams, such as the one shown in Figure 5, in module companion guides, is to provide a high-level, simplified view of module contents and flow so users understand the scope and main components of a module before diving into details.

Figure 5: Sample Spina Bifida Visio Diagram



APPENDIX B: PROJECT RESOURCES

The resources below provide easy access to project resources discussed in this report as well as detailed information about Synthea and links to additional PCOR data infrastructure information and projects.

ONC Synthetic Health Data Generation to Accelerate PCOR Project

- Project Webpage: https://www.healthit.gov/topic/scientific-initiatives/pcor/synthetic-health-data-generation-accelerate-patient-centered-outcomes
- Synthetic Health Data Frequently Asked Questions (FAQ): https://www.healthit.gov/sites/default/files/page/2021-10/Synthetic%20Health%20Data%20Project%20FAQs_508.pdf

Newly Developed Synthea Modules and Module Resources

- Prescribing Opioids for Chronic Pain and Treatment of Opioid Use Disorder
  - Companion Guide
  - Module in Synthea
  - Treatment of Sialorrhea in Cerebral Palsy Companion Guide
  - Module in Synthea
- Sepsis
  - Companion Guide
  - Module in Synthea
- Spina Bifida
  - Companion Guide
  - Module in Synthea
- Acute Myeloid Leukemia
  - Companion Guide
  - Module in Synthea

Synthetic Health Data Challenge

- Synthetic Health Data Challenge webpage (informational only): https://www.challenge.gov/challenge/synthetic-health-data-challenge/
- Synthetic Health Data Challenge Winning Solutions Webinar: https://youtu.be/gNBv5TFHRac?t=1

About Synthea

- Synthea wiki: https://github.com/synthetichealth/synthea/wiki
- Synthea Technical Guidance and Tips: https://www.healthit.gov/sites/default/files/page/2021-10/Synthetic%20Health%20Data%20Challenge_Technical%20Guidance%20and%20Tips_508.pdf

PCOR Projects and the PCOR Trust Fund

- ONC PCOR Projects: https://www.healthit.gov/topic/scientific-initiatives/building-data-infrastructure-support-patient-centered-outcomes-research
- About the Office of the Secretary of Health and Human Services Patient Centered Research Trust Fund (OS-PCORTF): https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf/about-os-pcortf

References

1. Synthetic Health Data Generation to Accelerate Patient-Centered Outcomes Research. Available from https://www.healthit.gov/topic/scientific-initiatives/pcor/synthetic-health-data-generation-accelerate-patient-centered-outcomes. Accessed Oct 2021.
2. HealthIT.gov: About ONC. Available from: https://www.healthit.gov/topic/about-onc. Accessed Jan 2022.
3. Chido, J. Simulated patient data fuels a new tool for healthcare innovators. The MITRE Corporation. Sep 2017. Available from: https://www.mitre.org/publications/project-stories/simulated-patient-data-fuels-a-new-tool-for-healthcare-innovators. Accessed Jan 2022.
4. The Office of the Assistant Secretary for Planning and Evaluation. Patient-Centered Outcomes Research Trust Fund [Internet]. Washington D.C.: ASPE 2020 [cited 2020 Mar 11]. Available from: https://aspe.hhs.gov/patient-centered-outcomes-research-trust-fund. Accessed Oct 2021.
5. HealthIT.gov: The United States Core Data for Interoperability (USCDI). Available from: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi Accessed Jan 2022.
6. HealthIT.gov: Health IT for Pediatric Care and Practice Settings. Available from: https://www.healthit.gov/topic/health-it-pediatric-care-and-practice-settings Accessed Jan 2022.
7. HealthIT.gov: Health IT Playbook, Section 4: Opioid Epidemic & Health IT. Available from: https://www.healthit.gov/playbook/opioid-epidemic-and-health-it/ Accessed Jan 2022.
8. Synthea Generic Module Framework, Examplitis Walk Through. Available from: https://github.com/synthetichealth/synthea/wiki/Generic-Module-Framework:-Complete-Example#examplitis-walk-through. Accessed Jun 2021.
9. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one. 2011;6(12): e28071.
10. Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name (a re-identification experiment). arXiv preprint. 2013:1304.7605.
11. Corporate Overview | The MITRE Corporation. Available from: https://www.mitre.org/about/corporate-overview. Accessed January 2022.
12. Synthea Wiki, Default Demographic Data. Available from: https://github.com/synthetichealth/synthea/wiki/Default-Demographic-Data. Accessed June 2021
13. Synthea Wiki, Getting Started. Available from: https://github.com/synthetichealth/synthea/wiki/Getting-Started. Accessed June 2021.
14. Care Pathways | AACPDM - American Academy for Cerebral Palsy and Developmental Medicine [Internet]. [cited 2020 May 20]. Available from: https://www.aacpdm.org/publications/care-pathways
15. Synthea Module Builder. Available from: https://synthetichealth.github.io/module-builder/. Accessed June 2021.
16. Synthea GitHub. Available from: https://github.com/synthetichealth/synthea. Accessed September 2021.

17. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inform Decis Mak. 2019 Mar 14;19(1):44.

18. Bozkurt M, Harman M. Automatically generating realistic test input from web services. Service Oriented System Engineering (SOSE), IEEE 6th International Symposium on Service Oriented System Engineering. Dec 2011:13-24. doi:10.1109/SOSE.2011.6139088.

19. Whiting M, Haack J, Varley C. Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. Proceedings of the 2008 Workshop on Beyond Time and Errors: Novel evaluation methods for Information Visualization. 2008.

20. Williams KJH, Ford RM, Bishop I, Loiterton ID, et al. Realism and selectivity in data-driven-visualizations: A process for developing viewer-oriented landscape surrogates. Landscape and Urban Planning. 2007: 81 (3), 213-224. https://doi.org/10.1016/j.landurbplan.2006.11.008

21. McLachlan S, Dube K, Gallagher T. Managing Realism in Synthetic Data Generation. Manuscript submitted to JAMIA. 2017: 1-141.

22. CDC Guideline for Prescribing Opioids for Chronic Pain —United States, 2016. MMWR RecommRep [Internet]. 2016; 65. Available from: https://www.cdc.gov/mmwr/volumes/65/rr/rr6501e1.htm

23. American Society of Addiction Medicine (ASAM). ASAM National Practice Guideline for the Use of Medications in the Treatment of Addiction Involving Opioid Use. 2015. Available from: https://newmexico.networkofcare.org/content/client/1446/2.6_17_AsamNationalPracticeGuidelines.pdf. Accessed Jan 2022.

24. Hotchkiss RS, Karl IE. The pathophysiology and treatment of sepsis. N Engl J Med. 2003 Jan 9;348(2):138–50.

25. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Sepsis Shock:2016. Critical Care Medicine [Internet] [cited 2020 Oct 16]. Available from: https://journals.lww.com/ccmjournal/Fulltext/2017/03000/Surviving_Sepsis_Campaign___International.15.aspx.

26. What is Spina Bifida? Resources and Prevention [Internet]. Spina Bifida Association. [cited 2020 Oct 18]. Available from: https://www.spinabifidaassociation.org/what-is-spina-bifida-2/

27. Care Coordination Guideline [Internet]. Spina Bifida Association. [cited 2020 Jul 1]. Available from: https://www.spinabifidaassociation.org/resource/care-coordination/

28. CDC. Spina Bifida Data and Statistics | CDC [Internet]. Centers for Disease Control and Prevention. 2019 [cited 2020 Aug 2]. Available from: https://www.cdc.gov/ncbddd/spinabifida/data.html

29. McCormick M, Friehling E, Kalpatthi R, Siripong N, et al. Cost-effectiveness of levofloxacin prophylaxis against bacterial infection in pediatric patients with acute myeloid leukemia. Pediatric Blood & Cancer. 2020 Oct;67(10):e28469.

30. Synthea Module Gallery. Available from: https://github.com/synthetichealth/synthea/wiki/Module-Gallery. Accessed Jun 2021.

31. HHS.gov. HHS updates interoperability standards to support the electronic exchange of sexual orientation, gender identity and social determinants of health. Jul 2021. Available from: https://www.hhs.gov/about/news/2021/07/09/hhs-updates-interoperability-standards-to-support-electronic-exchange-of-sogi-sdoh.html. Accessed Jan 2022.